Extracting Multimodal Learngene in CLIP: Unveiling the Multimodal Generalizable Knowledge

Ruiming Chen, Junming Yang, Shiyu Xia, Xu Yang*, Xin Geng *

School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
{220232251, jmingyang, shiyu_xia, xuyang_palm, xgeng}@seu.edu.cn

Abstract

CLIP (Contrastive Language-Image Pre-training) has attracted widespread attention for its multimodal generalizable knowledge, which is significant for downstream tasks. However, the computational overhead of a large number of parameters and large-scale pre-training poses challenges of pre-training a different scale of CLIP. *Learngene* extracts the generalizable components termed as *learngene* from an ancestry model and initializes diverse descendant models with it. Previous Learngene paradigms fail to handle the generalizable knowledge in multimodal scenarios. In this paper, we put forward the idea of utilizing a multimodal block to extract the multimodal generalizable knowledge, which inspires us to propose **MM-LG** (**M**ulti**m**odal **L**earn**g**ene), a novel framework designed to extract and leverage generalizable components from CLIP. Specifically, we first establish multimodal and unimodal blocks to extract the multimodal and unimodal generalizable knowledge in a weighted-sum manner. Subsequently, we employ these components to numerically initialize descendant models of varying scales and modalities. Extensive experiments demonstrate MM-LG's effectiveness, which achieves performance gains over existing learngene approaches (*e.g.*,+3.1% on Oxford-IIIT PET and +4.13% on Flickr30k) and comparable or superior results to the pre-training and fine-tuning paradigm (*e.g.*,+1.9% on Oxford-IIIT PET and +3.65% on Flickr30k). Notably, MM-LG requires only around 25% of the parameter storage while reducing around 2.8× pre-training costs for diverse model scales compared to the pre-training and fine-tuning paradigm, making it particularly suitable for efficient deployment across diverse downstream tasks.

Introduction

As a representative paradigm in multimodal research, CLIP (Contrastive Language-Image Pre-training) (Radford et al. 2021) leverages 400 million large-scale image-text pairs for contrastive pre-training. It achieves powerful representation capabilities and effectively integrates both modalities, demonstrating strong cross-modal generalization ability. These characteristics are manifested in excelling in various downstream tasks, including zero-shot image classification (Novack et al. 2023), cross-modal retrieval (Lülf et al. 2024; Cao et al. 2022, 2025; Xu et al. 2025) and image captioning (Mokady and Hertz 2021). Despite these advancements, an excessive number of parameters to be deployed has become a barrier for edge devices (*e.g.* Raspberry Pi, NVIDIA Jetson) (Yang et al. 2024b,a). Besides, if a different model scale is required, the repetitive pre-training is extremely time-consuming and computationally expensive. It is significant to extract the generalizable knowledge in CLIP to perform an efficient initialization for CLIP-based models.

Learngene, first proposed by (Wang et al. 2022), preserves the generalizable component of the pre-trained model and initializing diverse downstream models with it, as depicted in Figure. 1(a). The first stage is to extract the generalizable component, termed as **learngene**, from a large pre-trained model (referred to as the ancestry model). In the

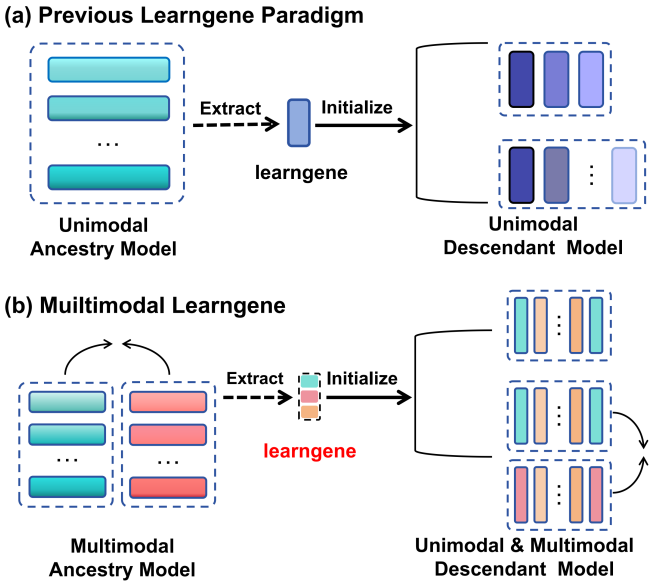

Figure 1: (a) Learngene paradigm comprises two stages. The first stage is to extract the compact learngene from an ancestry model. Secondly, diverse descendant models are initialized with it. (b) Multimodal Learngene first extracts learngene, which is generalizable across modalities, and subsequently initializes multimodal and unimodal descendant models of varying scales with it.

*Co-corresponding auther.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

second stage, this extracted learngene is utilized to initialize diverse models (referred to as the descendant model) for downstream tasks. Our proposed Learngene paradigm focuses on multimodal generalizable knowledge extraction.

Recent studies have extensively investigated and validated the effectiveness of the Learngene paradigm (Wang et al. 2023; Xia et al. 2024b,a; Feng et al. 2024; Lin et al. 2024, 2025). He-LG (Wang et al. 2022) proposed a gradient-based approach to extract learngene, which is used to initialize descendant models stacked with randomly initialized layers. TLEG (Xia et al. 2024a) developed a linear extraction method specifically for Transformer-based ancestry models, employing initialization techniques for descendant models. However, for these existing works, there is a lack of research into leveraging multimodal generalizable knowledge for learngene extraction within multimodal models, which is significant for broadening Learngene research.

To advance the exploration of generalizable learngene in multimodal architectures, we conduct our investigation on CLIP. As shown in Figure. 1 (b), to preserve the multimodal generalizable knowledge in CLIP explicitly, we seek to construct a multimodal block to handle it. Since there is unimodal generalizable knowledge in CLIP, unimodal blocks are established to store these components to assist the multimodal block, represented as vision and language blocks. Furthermore, to effectively extract generalizable knowledge into these blocks from CLIP, an auxiliary model (Xia et al. 2024a) must be constructed, wherein the parameters of each layer are the weighted sum of the multimodal block and the unimodal block with learnable coefficients.

Based on the analysis above, we propose a novel framework for multimodal learngene extraction from CLIP, termed as **M**ulti**m**odal **L**earn**g**ene (**MM-LG**). As illustrated in Figure. 3, MM-LG comprises two distinct stages: 1) **Extraction**: We construct an auxiliary model structure to extract the learngene from the pre-trained CLIP with the distillation technique (Hinton, Vinyals, and Dean 2015; Wang et al. 2024b). The extracted learngene is composed of blocks for multimodal, vision and language respectively, and coefficients. They are arranged in a weighted-sum manner within the extraction. 2) **Initialization**: We employ the extracted learngene to obtain the initial parameter values each layer for descendant models of diverse modalities and scales by integrating the coefficients and the blocks in a weighted-sum manner. With generalizable knowledge among modalities explicitly extracted, MM-LG effectively extracts the generalizable components of CLIP and initializes the descendant models of diverse scales to handle various multimodal or unimodal downstream tasks.

Our learngene extraction technique enables descendant models to achieve superior generalization capabilities across both multimodal and unimodal tasks at various model scales. To validate its effectiveness, we conduct comprehensive experiments across three distinct tasks: image classification, cross-modal retrieval, and image captioning. 1) In comparison with existing learngene approaches, our method demonstrates notable improvements, achieving performance gains of 3.1% on Oxford-IIIT PET (Parkhi et al. 2012) for the 8-layer configuration (55.07M) and 4.13% on Flickr30k (Young et al. 2014) for the 12-layer configuration (69.20M) compared to TLEG (Xia et al. 2024a). 2) Our approach achieves comparable and sometimes superior performance to the upper bound PT-FT (Pre-training and Fine-tuning) baseline, surpassing it by 1.9% on Oxford-IIIT PET and 3.65% on Flickr30k in the 12-layer setting (69.20M). 3) Our method requires only around 25% of the parameter storage of the PT-FT paradigm while enabling flexible model initialization, which reduces around 2.8$\times$ pre-training costs across different model scales.

These experimental results demonstrate our method's effectiveness in extracting generalizable components from CLIP that can be successfully applied across diverse downstream tasks.

Our main contributions are summarized as follows:

- Our work is the first to realize the significance of multimodal generalizable knowledge and explore it, which has not been considered in previous Learngene studies.
- We introduce MM-LG, a novel framework for extracting and expanding multimodal learngene from CLIP, which captures both multimodal and unimodal generalizable knowledge from CLIP.
- Extensive experiments show that MM-LG outperforms existing learngene methods and achieves comparable or superior results to the PT-FT paradigm across diverse tasks storing only around 25% of the parameters without repetitive pre-training.

Related Work

CLIP

Some previous multimodal works have investigated methods for establishing meaningful interactions between visual and linguistic modalities (Li et al. 2023; Jia et al. 2021; Qiao et al. 2024; Peng et al. 2025; Wu et al. 2025). CLIP (Radford et al. 2021) pioneered the establishment of representational relationships between images and text through contrastive learning. Subsequently, some recent studies (Lai et al. 2023; Zhang et al. 2024; Jiang et al. 2025; Li et al. 2025) have delved further into extracting cross-modal commonalities within the framework of CLIP. ALBEF (Li et al. 2021) proposed cross-modal attention to align the image and text representations with the fusion module. CLIP-CID (Yang et al. 2024b) adapted a cluster instance method to find commonalities in multimodal data. CLIP-KD (Yang et al. 2024a) explored several techniques to enhance the performance and efficiency of lightweight CLIP models through knowledge distillation. Clipping (Pei et al. 2023) proposed a layerwise alignment method to effectively transfer the knowledge from a large pre-trained model to a smaller model. However, despite the efforts of these methods in exploring the path to model compression and performance improvement for CLIP models, they face a significant challenge: when the scale of downstream-task models varies, these methods invariably require repetitive large-scale pre-training. Compared with other works for CLIP, our work aims to explore multimodal generalization ability, extracting the generalizable components in CLIP and utilizing them to initialize diverse models for downstream tasks.

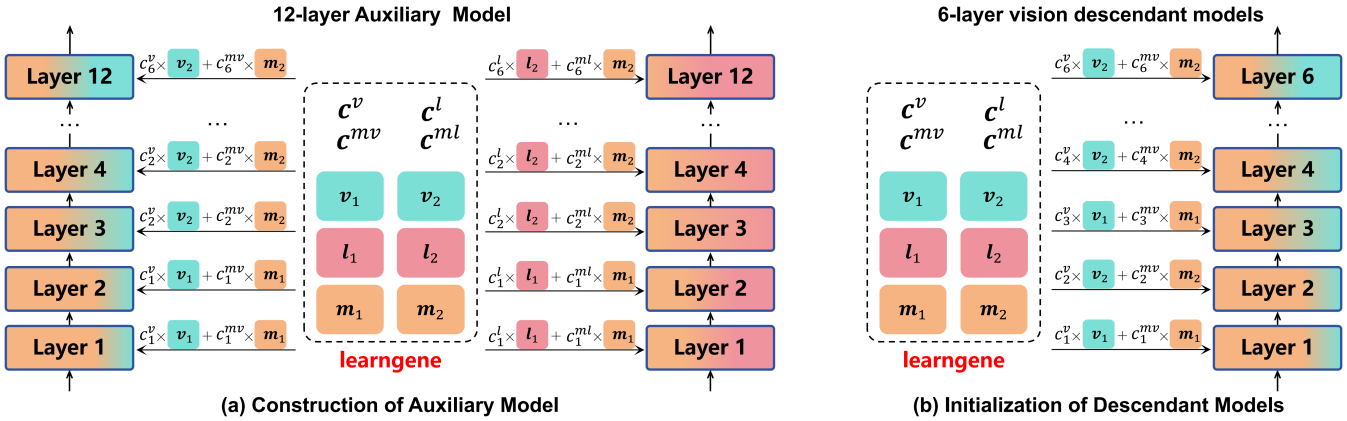

Figure 2: (a) Take a 12-layer auxiliary model for instance, during its construction, two groups of blocks $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ are alternately and repeatedly utilized. Each group is shared twice before switching to the other. For each vector in $\boldsymbol{c}_{LG}$, every scalar weight element is shared twice before proceeding to the next one. (b) With the learngene group index and coefficient index selection in the 12-layer auxiliary model eliminating repetitions, with well-extracted learngene, the modified index selection would serve as a parameter-value initialization for a normal Transformer-based 6-layer vision descendant model.

Learngene

Learngene presents an innovative and effective approach for extracting a compact yet information-rich component, termed as *learngene*, from a well-trained large-scale model, known as the Ancestry Model (Ans-Net). This learngene is then utilized to initialize Descendant Models (Des-Nets) of varying sizes. Several distinct methodologies have been developed to implement this approach (Wang et al. 2024a, 2023; Feng, Wang, and Geng 2024; Xie et al. 2024; Lin et al. 2024, 2025). He-LG (Wang et al. 2022) initially proposed extracting higher-level network layers as the learngene and integrating them with randomly initialized layers to construct Des-Nets. TLEG (Xia et al. 2024a) introduced a method for learngene extraction and expansion using structures with linear constraints. LearngenePool (Shi et al. 2024) employs a diverse strategy by extracting multiple small models from the Ans-Net to form learngene instances, which are then combined to create Des-Nets. SWS (Xia et al. 2024b) constructs an auxiliary model with stage-wise weight sharing to learn the learngene, which initializes Des-Nets of varying sizes. Compared with previous methods, our paradigm explores the multimodal generalizable knowledge across modalities in CLIP, which is significant for initializing descendant models of varying sizes for both unimodal and multimodal tasks.

Approach

Figure. 3 illustrates the pipeline of our proposed method, which consists of two distinct stages. In the first stage, our focus lies in extracting learngene from the ancestry model with the auxiliary model. The auxiliary model is structured with two groups of blocks and a series of coefficients, and notably, these are referred to as learngene. For the second stage, having successfully obtained the well-extracted learngene from stage 1, we leverage it to obtain initial parameter values for descendant models of varying sizes and diverse modalities. Subsequently, we would delve into the details.

Construction of Auxiliary Model

We would delve into how to construct the auxiliary model to extract multimodal learngene in CLIP. Given that CLIP is a multimodal model with dual architecture, it follows that the components of the learngene ought to incorporate at least these two modalities. Additionally, in CLIP, there is cross-modal interaction during contrastive learning based on the outputs of these two encoders. Consequently, an additional multimodal part will be incorporated into our learngene. To be more specific, our leargene consists of a block part $\boldsymbol{\theta}_{LG}$ and a coefficient part $\boldsymbol{c}_{LG}$, which are shared across the auxiliary model. Here for model structure optimization and facilitate harmonious interactions, $\boldsymbol{\theta}_{LG}$ are divided into 2 groups $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, where $\boldsymbol{\theta}_1 = \{\boldsymbol{\theta}_1^v, \boldsymbol{\theta}_1^l, \boldsymbol{\theta}_1^m\}$ and $\boldsymbol{\theta}_2 = \{\boldsymbol{\theta}_2^v, \boldsymbol{\theta}_2^l, \boldsymbol{\theta}_2^m\}$. Here $\boldsymbol{\theta}^v, \boldsymbol{\theta}^l, \boldsymbol{\theta}^m$ refer to the blocks for vision, language and multimodal respectively and the subscript is the group index. Each $\boldsymbol{\theta}$ represents the entirety of weights and biases of linear processes for multi-head self-attention (MSA) and multi-layer perceptron (MLP) in a Transformer layer (Vaswani et al. 2017). $\boldsymbol{c_{LG}}$ consists of 3 parts $\boldsymbol{c}^v, \boldsymbol{c}^l$ and $\boldsymbol{c}^m$, where $\boldsymbol{c}^m = \{\boldsymbol{c}^{mv}, \boldsymbol{c}^{ml}\}$. $\boldsymbol{c}^v, \boldsymbol{c}^l, \boldsymbol{c}^{mv}, \boldsymbol{c}^{ml}$ refer to the coefficient vectors for vision, language, multimodal-vision, and multimodal-language respectively with half the length of the layer sequence.

Subsequently, we proceed to the construction part of the auxiliary model, the details of which are illustrated in Figure. 2(a). During the process of constructing an auxiliary model layer, two groups of blocks $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ are used alternately and repeatedly. That is, one group is shared twice and then switched to the other, with them taking turns in this way. For each vector in $\boldsymbol{c}_{LG}$, with its length half the number of layers, every weight scalar element is shared twice prior to the next one. Each time the group of learngene blocks and the weight scalars are confirmed, a Transformer layer is con-

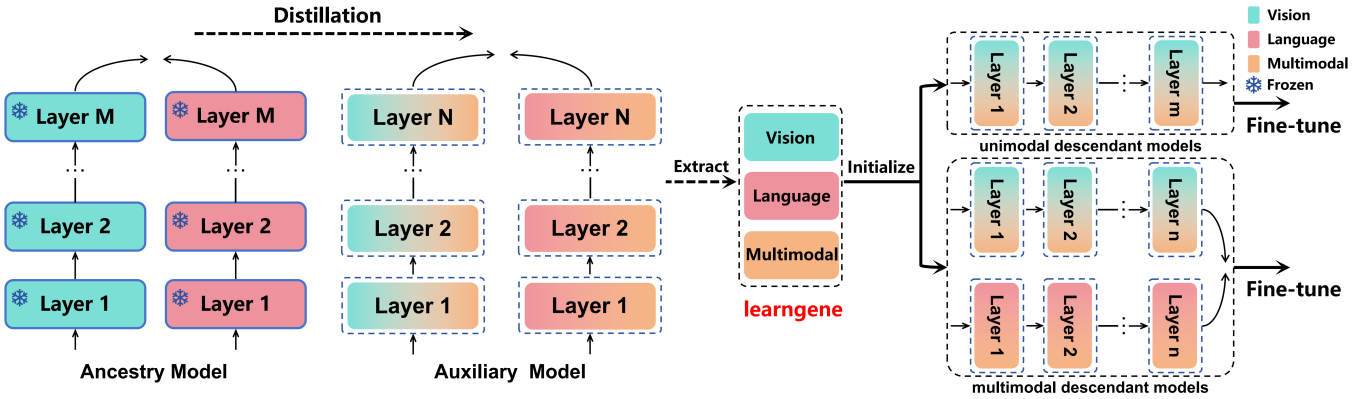

Figure 3: The framework of our proposed method comprises two stages. In the first stage, we construct an auxiliary model wherein the parameters of each layer are the weighted sum of a multimodal block and a unimodal block with learnable coefficients, and we subsequently train it through distillation against the ancestry model. After obtaining learngene composed of blocks and coefficients, in the second stage, we compute the initial parameter values for both unimodal and multimodal descendant models of varying depths, which are fine-tuned for downstream tasks.

structed. For layer i in the vision encoder, the block group index j and the coefficient index k could be worked out with i and the mentioned arrangement. Considering parameters $\boldsymbol{P}_i^v$ in Transformer layer i, we have Eq. 1.

$$\boldsymbol{P}_i^v = c_k^v \times \boldsymbol{\theta}_j^v + c_k^{mv} \times \boldsymbol{\theta}_j^m \tag{1}$$

For layer i in the language encoder, we similarly have Eq. 2.

$$\boldsymbol{P}_i^l = c_k^l \times \boldsymbol{\theta}_j^l + c_k^{ml} \times \boldsymbol{\theta}_j^m \tag{2}$$

Take layer 6 for instance, the block group index is 2 and the coefficient index is 3. Thus $\boldsymbol{P}_6^v$ for vision branch is $\boldsymbol{P}_6^v = c_3^v \times \boldsymbol{\theta}_2^v + c_3^{mv} \times \boldsymbol{\theta}_2^m$ while for language branch is $\boldsymbol{P}_6^l = c_3^l \times \boldsymbol{\theta}_2^l + c_3^{ml} \times \boldsymbol{\theta}_2^m$.

Except for the parameters referred above, other components like Layer Normalization(LN) (Ba, Kiros, and Hinton 2016) part or Embedding part are not decomposed. The LN part is shared in each Transformer layer.

Training of Auxiliary Model

To extract the learngene in CLIP, we take advantage of knowledge distillation (Hinton, Vinyals, and Dean 2015) to train the auxiliary model. Combined with the learngene and components other than the learngene, a dual architecture auxiliary model can already be established. Subsequently, the output of this model can be exploited to conduct distillation on CLIP.

Before discussing the details, we abbreviate the auxiliary model as aux and the ancestry model as anc. Assume $\{(I_k, T_k)\}_{k=1}^{|\mathcal{B}|}$ is a mini-batch of image-text pair data, with the vision and language encoders in the auxiliary model $f_v^{aux}(\cdot)$ and $f_l^{aux}(\cdot)$, d-dimension features $v_k^{aux} = f_v^{aux}(I_k)$, $s_k^{aux} = f_l^{aux}(T_k)$ are obtained. Subsequently, $l2$ normalization is applied for all features. Following the implementation in (Radford et al. 2021), we firstly obtain the pairwise cosine similarity logit $logit^{aux}$

$$logit^{aux} = V^{aux} S^{aux\top} / \tau^{aux} \tag{3}$$

where $V^{aux}, S^{aux} \in \mathbb{R}^{|\mathcal{B}| \times d}$ are feature batches for vision and language while τ^{aux} is a learnable temperature. Similarly, $logit^{anc}$ could be worked out where τ^{anc} is frozen. Associated with the Cross-Entropy function $CE(\cdot)$ and an identity matrix $\mathbf{I} \in \mathbb{R}^{|\mathcal{B}| \times |\mathcal{B}|}$, Contrastive Language-Image Pre-Training would be performed with Eq. 4

$$\mathcal{L}_{CLIP} = \frac{1}{2}(CE(logit^{aux}, \mathbf{I}) + CE(logit^{aux\top}, \mathbf{I})) \tag{4}$$

Except for $\mathcal{L}_{CLIP}$, the Cross-Entropy function is applied for the distillation part. With the differences calculated, the auxiliary model could effectively extract learngene from the ancestry model. The Cross-Entropy function is performed between two logits $logit^{aux}$ and $logit^{anc}$

$$\mathcal{L}_{dist}^{i2t} = CE(logit^{aux}, logit^{anc}) \tag{5}$$

$$\mathcal{L}_{dist}^{t2i} = CE(logit^{aux\top}, logit^{anc\top}) \tag{6}$$

Work out the average of them to get the distillation loss

$$\mathcal{L}_{dist} = \frac{1}{2}(\mathcal{L}_{dist}^{i2t} + \mathcal{L}_{dist}^{t2i}) \tag{7}$$

Additionally, to balance $\mathcal{L}_{CLIP}$ and $\mathcal{L}_{dist}$, a loss weight λ is used to integrate these two losses

$$\mathcal{L}_{train} = \mathcal{L}_{CLIP} + \lambda \mathcal{L}_{dist} \tag{8}$$

Initialization of Descendant Models

For this part, we employ the extracted learngene to make a initialization for parameter values in a unimodal Transformer-based model or a normal CLIP model, as shown in Figure. 2(b). We would discuss initializing a normal unimodal Transformer model and combining vision and language parts for a normal CLIP model.

To initialize a unimodal Transformer descendant model, components other than the learngene could be loaded directly like Embedding or LN. Similar to the process in auxiliary model construction, the key is to select the appropriate

Method	Layers	MSCOCO (Lin et al. 2014)		Flickr30k (Young et al. 2014)	
		I2T	T2I	I2T	T2I
PT-FT	12	30.56	30.10	61.53	59.86
Scratch		0.46	0.54	0.29	0.49
He-LG (Wang et al. 2022)		1.70	1.54	3.45	3.25
TLEG (Xia et al. 2024a)		29.28	26.56	61.05	57.89
MM-LG (Ours)		**33.06**	**31.48**	**65.18**	**65.18**
PT-FT	8	28.74	27.50	58.28	57.20
Scratch		0.80	0.90	1.28	0.49
He-LG (Wang et al. 2022)		1.46	1.36	2.66	3.06
TLEG (Xia et al. 2024a)		27.22	24.94	57.30	55.23
MM-LG (Ours)		**29.08**	**27.54**	**59.37**	**57.98**
PT-FT	6	**26.50**	**25.76**	**52.76**	**53.55**
Scratch		0.84	0.50	0.39	0.59
He-LG (Wang et al. 2022)		1.30	1.34	2.96	3.25
TLEG (Xia et al. 2024a)		24.02	24.40	47.83	49.11
MM-LG (Ours)		25.20	23.54	51.09	52.17

Table 1: Performance of cross-modal retrieval on COCO and Flickr30k datasets.

learngene group and coefficients to calculate the parameter values, which serves as a initialization for parameter values in normal models. The selection for the learngene group and coefficients is repetitive every 2 layers in Figure. 2(a). By eliminating such repetitions, with the modified index selection in Figure. 2(b), the result obtained by the weighted summation of each layer can serve as a initialization for parameter values in 6-layer descendant models. When constructing a descendant model of n layers, with the initialized 6-layer descendant model, we merely require to replicate each of the first $n-6$ layers once and then connect them in sequence to accomplish the initialization. Take a 8-layer normal Vision Transformer model (Dosovitskiy 2020) for instance, the first 2 layers of a 6-layer descendant model are replicated like those in the auxiliary model and connected with the remaining 4 layers to perform the 8-layer initialization.

Except for this parameter initialization, the LN part for each layer is initialized with the weight-shared one we have referred in Section for Construction of Auxiliary Model. After a minor activation (Muralidharan et al. 2024) with a small amount of data, the final model could be employed for further fine-tuning in downstream tasks. More details about the activation would be presented in Appendix.

Experiments

In this section, we mainly explore the following questions:

1. Whether MM-LG can handle multimodal downstream tasks with the extracted learngene.
2. Whether MM-LG can handle unimodal downstream tasks with the extracted learngene.
3. How MM-LG improves the efficiency.

Experimental Setup

Datasets. For the learngene extraction and the pre-training, we utilize Conceptual Captions 12M (CC12M) (Changpinyo et al. 2021) and Conceptual Captions 3M (CC3M) (Sharma et al. 2018). MSCOCO (Lin et al. 2014) and Flickr30k (Young et al. 2014) datasets are employed for cross-modal downstream tasks. As for visual tasks, we use CIFAR-100 (Krizhevsky, Hinton et al. 2009), Food-101 (Bossard, Guillaumin, and Van Gool 2014), and Oxford-IIIT PET (Parkhi et al. 2012) datasets.

Baselines. Scratch randomly initializes the weight and trains on the downstream dataset. The PT-FT paradigm (Pre-training and Fine-tuning) first conducts pre-training on large-scale datasets and then performs fine-tuning on downstream datasets. The distillation paradigm(Hinton, Vinyals, and Dean 2015) extends the PT-FT paradigm by integrating the pre-trained CLIP-ViT-B teacher model during pre-training. We extend the He-LG (Wang et al. 2022) method from unimodal to multimodal, and extract learngene from a distilled ViT-S CLIP to initialize the downstream models. Similarly, we extend TLEG (Xia et al. 2024a) to a multimodal model, impose linear constraints on each modality to extract learngene and expand it in downstream tasks.

Training details. We follow the consistent training details with CLIP-related works (Yang et al. 2024a; Ilharco et al. 2021; Cherti et al. 2023). Specifically, we extract a CLIP-ViT-S-sized auxiliary model from the pre-trained CLIP-ViT-B model, and then initialize descendant models following the requirements of downstream tasks. The experiments run over 8 Ascend 910B NPUs (64GB) hardware with RAM 1000GB. Please refer to Appendix for more details.

Cross-modal Retrieval

The results for the cross-modal retrieval task are shown in Table 1, and it is apparent that our paradigm MM-LG could handle this task with generalizable multimodal learngene.

Our framework has advantages under diverse layer settings and dataset settings compared to other learngene

Method	Layers	CIFAR100	Food101	PET
PT-FT	12	87.41	90.13	87.71
Scratch		71.08	75.76	32.71
He-LG		67.44	71.85	34.12
TLEG		86.34	89.24	87.14
MM-LG (Ours)		**87.70**	**90.15**	**89.60**
PT-FT	8	86.10	**89.65**	86.87
Scratch		66.81	73.39	30.55
He-LG		65.42	69.74	34.55
TLEG		85.15	88.59	84.34
MM-LG (Ours)		**86.24**	88.69	**87.39**
PT-FT	6	84.37	**88.44**	**84.04**
Scratch		67.78	73.64	29.03
He-LG		63.34	67.90	32.45
TLEG		83.79	86.75	82.24
MM-LG (Ours)		**84.50**	87.30	83.14

Table 2: Performance comparison across layers and methods on CIFAR-100, Food-101, and Oxford-IIIT PET datasets.

Method	Layers	BLEU@4↑	CIDEr↑	ROUGE-L↑
PT-FT	12	38.37	53.90	44.44
He-LG		24.36	14.72	34.70
TLEG		40.19	56.33	45.39
MM-LG (Ours)		**41.09**	**59.34**	**45.98**
PT-FT	8	36.81	47.61	42.97
He-LG		23.01	11.25	33.62
TLEG		36.43	45.63	42.52
MM-LG (Ours)		38.99	**53.28**	**44.74**
PT-FT	6	36.54	44.51	42.67
He-LG		22.86	13.07	33.54
TLEG		33.69	39.87	41.20
MM-LG (Ours)		38.16	**49.80**	**43.74**

Table 3: Performance of image captioning on COCO.

paradigms. Retrieval tasks are hard to handle without a sufficient initialization, thus downstream models trained from scratch manifest poor performance. He-LG also exhibits inferior performance since most layers in descendant models are directly trained without a sufficient initialization. In contrast, TLEG demonstrates normal performance but MM-LG presents substantially better performance than it. Take the image-to-text retrieval task for instance, MM-LG outperforms TLEG by **3.78%**, **1.86%** and **1.18%** respectively in the 12-layer, 8-layer and 6-layer settings on COCO, while **4.13%**, **2.07%** and **3.26%** in the same settings on Flickr30k. This result emphasizes the significance of constructing a multimodal block in learngene explicitly, which handles multimodal generalizable knowledge more effectively than employing linear expansion in both encoders.

Compared with the upper bound, the PT-FT paradigm, our framework surprisingly exceeds it by a margin of **2.50%**, **0.34%**, respectively in the 12-layer, 8-layer settings on COCO, while **3.65%**, **1.09%** in the same settings on Flickr30k. In the 6-layer setting, our paradigm achieves comparable results merely with a slight decrease since the PT-FT paradigm costs a lot for diverse model scales. The results for the distillation paradigm are listed in Appendix. Compared with these paradigms, MM-LG provides a superior initialization of generalizable multimodal knowledge without repetitive pre-training.

Image Classification

As illustrated in Table 2, merely applying part of the learngene, our method could handle unimodal tasks with vision and multimodal generalizable components.

Despite the traditional unimodal task of image classification, the performance demonstrated by different learngene paradigms varies substantially. For this task, the effect of negative transfer (Wang et al. 2019) occurs to He-LG. The trained parameters of the last three layers of the ancestry model actually impair the performance of the descendant models, whose performance is inferior to that of the models trained from scratch on CIFAR-100 and Food-101. Compared with TLEG, across all layer settings, MM-LG has an improvement of **0.7%**$\sim$ **1.4%** on CIFAR-100 and **0.1%**$\sim$ **1.0%** on Food-101. Even with the small-scale dataset Oxford-IIIT PET, MM-LG yet improves **2.46%**, **3.05%** and **0.90%** gains over TLEG. It is adequately demonstrated that MM-LG, while handling multimodal generalizable knowledge, effectively preserves unimodal generalizable knowledge. These two components handle downstream unimodal tasks remarkably well through their cooperation.

Compared with the PT-FT paradigm, MM-LG surpasses it by **0.29%**, **0.02%** and **1.89%** on the listed datasets in the 12-layer configuration and achieves comparable results in the other 2 layer settings. MM-LG even outperforms it by **0.52%** on Oxford-IIIT PET in the 8-layer setting. The results for the distillation paradigm are listed in Appendix. It is demonstrated that MM-LG has comprehensively extracted the generalizable knowledge across all modalities in CLIP.

Image Captioning

We conduct the image captioning task following (Mokady and Hertz 2021). After initializing the vision encoder, we connect it with a lightweight Transformer-based mapping network and the language model GPT-2 (Radford et al. 2019), subsequently training the mapping network with the vision encoder and GPT-2 frozen. Since no more fine-tuning is conducted for the vision part, models trained from scratch would not participate in the performance comparison.

The results are demonstrated in Table 3. Compared with other learngene paradigms, our method has achieved a significant advantage in this task. We apply BLEU (Papineni et al. 2001), CIDEr (Vedantam, Zitnick, and Parikh 2015) and ROUGE-L (Lin and Och 2004) for evaluation metrics. Taking BLEU for example, MM-LG has an improvement of **16.73**, **15.98**, **15.30** over He-LG in the 12-layer, 8-layer and 6-layer settings, and surpasses TLEG by **0.90**, **2.56**, **4.47** in the same settings. Our method also surpasses these

Method	Params(M)	GPU-hours(H)	Avg Acc(%)
PT-FT	151.4	1142.4	57.52
He-LG	**10.7**	459.6	3.02
TLEG	30.3	457.5	55.39
MM-LG (Ours)	37.4	**407.8**	**58.55**

Table 4: Performance of storage efficiency, training costs and the average image-to-text accuracy on Flickr30k.

paradigms for the other two matrices by a large margin in diverse layer settings. Although other learngene paradigms extract generalizable knowledge to some extent in CLIP, they are less comprehensive than MM-LG in exploring multimodal generalizable knowledge.

To our surprise, MM-LG outperforms the PT-FT paradigm in all layer settings in this task. Especially for the CIDEr metric, MM-LG significantly improves **5.44**, **5.67**, **5.29** gains over the PT-FT paradigm even though it has sufficient pre-training in all layer settings. The results for the distillation paradigm are listed in Appendix. Compared to the these paradigms, MM-LG acquires sufficient extraction before downstream tasks, which is more specific to multimodal and unimodal generalizable knowledge than repetitive pre-training with large-scale image-text pairs of data.

Efficiency

Table 4 presents the storage efficiency and the training costs before conducting downstream tasks and the average image-to-text performance on Flickr30k for the PT-FT paradigm and Learngene paradigms. Compared with the PT-FT paradigm, within the aspect of storage, our method merely requires 37.4M of parameter storage while the PT-FT paradigm requires 151.4M for three scales of models. We have achieved a reduction of approximately **75%** in storage. Furthermore, for training costs our paradigm demands 407.8 GPU-hours to accomplish the learngene extraction. In contrast, the PT-FT paradigm demands 1142.4 GPU-hours, nearly **2.8**$\times$ of ours because it demands a total pre-training for each model scale requirement.

Besides, MM-LG's average performance outperforms the PT-FT paradigm by **1.03%**. The results for the distillation paradigm are listed in Appendix. This metric significantly highlights the superiority of MM-LG in efficiency that this paradigm eliminates the need for repetitive pre-training for diverse model scales while effectively extracting the generalizable components in CLIP.

In comparison with the previous Learngene paradigms, the training costs are analogous to theirs since Learngene paradigms need merely a single extraction procedure before downstream tasks. The parameter storage for MM-LG is marginally larger due to the demand for preserving both multimodal and unimodal generalizable knowledge, leading to superior downstream performance.

Ablation Study

In the introduction section, it has been analyzed that there are a multimodal block, unimodal blocks and corresponding coefficients extracted in the first stage, representing the generalizable components for multimodal and unimodal modalities respectively. We conduct this ablation to verify whether the superior performance of MM-LG is due to one generalizable component alone or their collaboration.

Method	Layers	Flickr30k		CIFAR-100
		I2T	T2I	
w/o MM	12	23.47	22.88	78.11
only MM		23.37	22.98	81.15
MM-LG		**65.18**	**65.18**	**87.70**
w/o MM	8	18.34	19.23	78.15
only MM		18.44	19.13	78.81
MM-LG		**59.37**	**57.98**	**86.24**
w/o MM	6	13.12	13.12	75.11
only MM		12.92	13.81	76.79
MM-LG		**51.09**	**52.17**	**84.50**

Table 5: Ablation study on cross-modal retrieval Flickr30k dataset and image classification CIFAR-100 dataset.

The results are shown in Table 5, where "*w/o* MM" denotes initializing descendant models without the extracted multimodal part and "*only* MM" denotes initializing descendant models without the extracted unimodal part. For downstream multimodal tasks, we take image-to-text retrieval on Flickr30k for instance, MM-LG surpasses them by a large margin of about **42%**, **41%**, **38%** in the 12-layer, 8-layer, 6-layer settings accordingly, while for unimodal tasks, we outperform them by about **8%**, **8%**, **9%** on CIFAR-100 in the same layer settings. It is manifest that these two generalizable components are required to perform in concert, coherent with the cross-modal generalization ability in CLIP.

Conclusion

In this paper, inspired by employing a multimodal block to preserve multimodal generalizable knowledge in CLIP, we propose a novel and effective learngene extraction method, termed as MM-LG. MM-LG is capable of extracting the generalizable knowledge in both multimodal and unimodal modalities from the learngene extraction stage, initializing descendant models for diverse downstream scenarios with different depth and modality requirements. The experimental results verify that MM-LG surpasses other learngene paradigms and achieves comparable or superior performance to the pre-training and fine-tuning paradigm without repetitive pre-training.

Acknowledgments

This research was supported by the Jiangsu Science Foundation (BK20243012, BG2024036), the National Science Foundation of China (62125602, U24A20324, 92464301), and the Fundamental Research Funds for the Central Universities (2242025K30024). This work was partially supported by Southeast University Kunpeng&Ascend Center of Cultivation.

References

Ba, J.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. *ArXiv*, abs/1607.06450.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, 446–461. Springer.

Cao, M.; Li, S.; Li, J.; Nie, L.; and Zhang, M. 2022. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*.

Cao, M.; Zhou, X.; Jiang, D.; Du, B.; Ye, M.; and Zhang, M. 2025. Multilingual Text-to-Image Person Retrieval via Bidirectional Relation Reasoning and Aligning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3558–3568.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2818–2829.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, F.; Wang, J.; and Geng, X. 2024. Transferring Core Knowledge via Learngenes. *arXiv preprint arXiv:2401.08139*.

Feng, F.; Xie, Y.; Wang, J.; and Geng, X. 2024. Wave: Weight template for adaptive initialization of variable-sized models. *arXiv preprint arXiv:2406.17503*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jiang, L.; Zhang, Z.; Zeng, Y.; Xie, C.; Liu, T.; Li, Z.; Cheng, L.; and Xu, X. 2025. DCP: Dual-Cue Pruning for Efficient Large Vision-Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 21202–21215.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

Lai, Z.; Li, Z.; Oliveira, L. C.; Chauhan, J.; Dugger, B. N.; and Chuah, C.-N. 2023. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2374–2380.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, L.; Peng, Y.; Yang, X.; Cheng, R.; Xu, H.; Yan, M.; and Huang, F. 2025. L-clipscore: a lightweight embedding-based captioning metric for evaluating and training. *arXiv preprint arXiv:2507.08710*.

Lin, C.-Y.; and Och, F. J. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*.

Lin, S.; Yang, X.; Wang, Q.; Guo, S.; Kou, Z.; and Geng, X. 2025. ALPSB: Adaptive Learngene with Plastic and Stable Branches. *Pattern Recognition*, 112623.

Lin, S.; Zhang, M.; Chen, R.; Yang, X.; Wang, Q.; and Geng, X. 2024. Linearly decomposing and recomposing vision transformers for diverse-scale models. *Advances in Neural Information Processing Systems*, 37: 33188–33212.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lülf, C.; Lima Martins, D. M.; Vaz Salles, M. A.; Zhou, Y.; and Gieseke, F. 2024. CLIP-Branches: Interactive Fine-Tuning for Text-Image Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, 2719–2723. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704314.

Mokady, R.; and Hertz, A. 2021. ClipCap: CLIP Prefix for Image Captioning. *ArXiv*, abs/2111.09734.

Muralidharan, S.; Sreenivas, S. T.; Joshi, R. B.; Chochowski, M.; Patwary, M.; Shoeybi, M.; Catanzaro, B.; Kautz, J.; and

Molchanov, P. 2024. Compact language models via pruning and knowledge distillation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Novack, Z.; Mcauley, J.; Lipton, Z. C.; and Garg, S. 2023. CHiLS: Zero-Shot Image Classification with Hierarchical Label Sets. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 26342–26362. PMLR.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2001. BLEU. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Pei, R.; Liu, J.; Li, W.; Shao, B.; Xu, S.; Dai, P.; Lu, J.; and Yan, Y. 2023. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18983–18992.

Peng, Y.; Zhang, G.; Zhang, M.; You, Z.; Liu, J.; Zhu, Q.; Yang, K.; Xu, X.; Geng, X.; and Yang, X. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Qiao, Y.; Duan, H.; Fang, X.; Yang, J.; Chen, L.; Zhang, S.; Wang, J.; Lin, D.; and Chen, K. 2024. Prism: A framework for decoupling and assessing the capabilities of vlms. *Advances in Neural Information Processing Systems*, 37: 111863–111898.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Shi, B.; Xia, S.; Yang, X.; Chen, H.; Kou, Z.; and Geng, X. 2024. Building Variable-Sized Models via Learngene Pool. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14946–14954.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *Neural Information Processing Systems,Neural Information Processing Systems*.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based Image Description Evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Q.; Yang, X.; Chen, H.; and Geng, X. 2024a. Vision Transformers as Probabilistic Expansion from Learngene. In *Forty-first International Conference on Machine Learning*.

Wang, Q.; Yang, X.; Lin, S.; Wang, J.; and Geng, X. 2023. Learngene: Inheriting condensed knowledge from the ancestry model to descendant models. *arXiv preprint arXiv:2305.02279*.

Wang, Q.-F.; Geng, X.; Lin, S.-X.; Xia, S.-Y.; Qi, L.; and Xu, N. 2022. Learngene: From open-world to your learning task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8557–8565.

Wang, Y.; Cheng, L.; Duan, M.; Wang, Y.; Feng, Z.; and Kong, S. 2024b. Improving knowledge distillation via regularizing feature direction and norm. In *European Conference on Computer Vision*, 20–37. Springer.

Wang, Z.; Dai, Z.; Poczos, B.; and Carbonell, J. 2019. Characterizing and Avoiding Negative Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Y.; Zhou, Y.; Ziheng, Z.; Peng, Y.; Ye, X.; Hu, X.; Zhu, W.; Qi, L.; Yang, M.-H.; and Yang, X. 2025. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*.

Xia, S.; Zhang, M.; Yang, X.; Chen, R.; Chen, H.; and Geng, X. 2024a. Transformer as Linear Expansion of Learngene. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16014–16022.

Xia, S.-Y.; Zhu, W.; Yang, X.; and Geng, X. 2024b. Exploring Learngene via Stage-wise Weight Sharing for Initializing Variable-sized Models. *arXiv preprint arXiv:2404.16897*.

Xie, Y.; Feng, F.; Wang, J.; Geng, X.; and Rui, Y. 2024. Kind: Knowledge integration and diversion in diffusion models. *arXiv preprint arXiv:2408.07337*.

Xu, Y.; Wu, M.; Guo, Z.; Cao, M.; Ye, M.; and Laaksonen, J. 2025. Efficient text-to-video retrieval via multi-modal multi-tagger derived pre-screening. *Visual Intelligence*, 3(1): 1–13.

Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024a. CLIP-KD: An Empirical Study of CLIP Model Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15952–15962.

Yang, K.; Gu, T.; An, X.; Jiang, H.; Dai, X.; Feng, Z.; Cai, W.; and Deng, J. 2024b. CLIP-CID: Efficient CLIP Distillation via Cluster-Instance Discrimination. *ArXiv*, abs/2408.09441.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zhang, D.; Yang, J.; Lyu, H.; Jin, Z.; Yao, Y.; Chen, M.; and Luo, J. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*.