Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

Prompt learning in computer vision: a survey*

Yiming LEI[†‡1], Jingqi LI[1], Zilong LI[1], Yuan CAO[1], Hongming SHAN[†‡2,3,4]

[1]*Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200438, China*

[2]*Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China*

[3]*MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200433, China*

[4]*Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai 201210, China*

[†]E-mail: ymlei@fudan.edu.cn; hmshan@fudan.edu.cn

Received May 31, 2023; Revision accepted Oct. 17, 2023; Crosschecked Jan. 5, 2024

Abstract: Prompt learning has attracted broad attention in computer vision since the large pre-trained vision-language models (VLMs) exploded. Based on the close relationship between vision and language information built by VLM, prompt learning becomes a crucial technique in many important applications such as artificial intelligence generated content (AIGC). In this survey, we provide a progressive and comprehensive review of visual prompt learning as related to AIGC. We begin by introducing VLM, the foundation of visual prompt learning. Then, we review the vision prompt learning methods and prompt-guided generative models, and discuss how to improve the efficiency of adapting AIGC models to specific downstream tasks. Finally, we provide some promising research directions concerning prompt learning.

Key words: Prompt learning; Visual prompt tuning (VPT); Image generation; Image classification; Artificial intelligence generated content (AIGC)

https://doi.org/10.1631/FITEE.2300389 **CLC number:** TP181

1 Introduction

Prompt learning was first proposed in the natural language processing (NLP) community, which endows language models with the ability to model raw texts directly (Lu YN et al., 2022). Specific research directions include pre-trained models, prompt engineering, and prompt fine-tuning. Based on the great success of prompt learning, it is noteworthy to discuss how prompt learning improves computer vision (CV) tasks, which can inspire more valuable research and vision applications. As shown in Fig. 1, for a wide range of CV tasks, prompt embedding, yielded by traditional prompt and prompt learning, is a direct controlling condition that endows outputs of generative models with animated contents and improves the performance of discriminative models.

For the past decade, vision tasks have been largely enhanced by convolutional neural networks (CNNs) and corresponding techniques such as attention mechanism and skip connection. Various CNN variants have progressively improved the performances of tasks including image classification, object detection, and semantic segmentation. Because the Transformer has empowered NLP models with generalized representations, the vision Transformer (ViT) has triggered a new revolution in the CV community (Dosovitskiy et al., 2021; Khan et al., 2022; Han et al., 2023; Shamshad et al., 2023). For the

‡ Corresponding authors

* Project supported by the National Natural Science Foundation of China (Nos. 62306075 and 62101136), the China Postdoctoral Science Foundation (No. 2022TQ0069), the Natural Science Foundation of Shanghai, China (No. 21ZR1403600), the Shanghai Municipal of Science and Technology Project, China (No. 20JC1419500), and the Shanghai Center for Brain Science and Brain-Inspired Technology, China

ORCID: Yiming LEI, https://orcid.org/0000-0002-1349-7074; Hongming SHAN, https://orcid.org/0000-0002-0604-3197

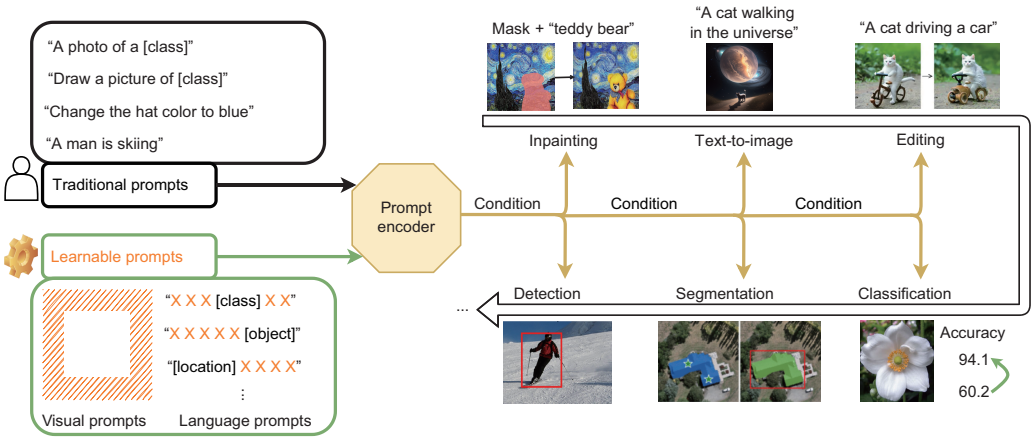

Fig. 1 Pipeline of learning with traditional and learnable prompts. Prompt learning is a critical technique in current AIGC tasks. Apart from human-understandable prompts, learnable prompts have been proposed to further enhance specific tasks

vision tasks, current ViT-based networks have surpassed the corresponding CNN counterparts with a relatively equal magnitude of trainable parameters. In addition to these high-level recognition tasks, ViT works well in low-level image processing tasks, such as image denoising, deblurring, and deraining.

Because the Transformer has been successful in both CV and NLP fields, the large-scale vision-language models (VLMs) pre-trained on large numbers of image-text pairs, have subverted the learning paradigms of many visual tasks and serve as the foundation of prompt learning in CV. A typical framework is contrastive language-image pre-training (CLIP), which aligns visual and textual features through contrastive learning (Radford et al., 2021). Compared with CLIP, A Large-scale ImaGe and Noisy-text embedding (ALIGN) conducts the same contrastive learning scheme and pre-trains the image and text encoders using more training data with noisy samples (Jia C et al., 2021). This kind of pre-trained model formulates the image classification problem as an image-text matching scheme, enriching the learning information against the traditional one-hot approximation. Notably, these methods have amazing few-shot, zero-shot learning, and image-text retrieval capabilities, and hence, CLIP provides follow-up research with fundamental learning paradigms and powerful image and text encoders.

The success of CLIP is inseparable from the prompt engineering and prompt ensemble, which help learn generalized multi-modal representations. In the NLP community, prompt-based learning enables language models to predict the probability of raw text directly, where the prompt refers to a language template describing the corresponding category information (Liu PF et al., 2023). This framework allows successful pre-training of large-scale language models (LLMs) using massive raw texts; for example, the texts are crawled from the Internet directly. Although the effectiveness of applying prompt engineering for VLMs has been validated in CLIP, only hand-crafted language templates are used, which hinders further performance improvement for downstream tasks, such as "a photo of a [CLS]" and "a blurry photo of the [CLS]," where "[CLS]" denotes the class token. To achieve more generalized prompts for facilitating downstream image classification, Zhou KY et al. (2022a, 2022b) proposed to learn trainable prompts through contrastive learning. Then, many studies on language prompting emerged and intended to learn more explainable text tokens correlated with image features. In addition, inspired by such language-driven prompt learning, visual prompting aims at adapting large models to downstream tasks depending on the constructed image prompt inputs, which is different from the aims of language prompting (Bahng et al., 2022; Bar et al., 2022; Xing et al., 2022; Chen AC et al., 2023; Zhu JW et al., 2023).

Based on the efficient prompt learning paradigms mentioned above, language-driven prompting facilitates traditional deep generative models like generative adversarial network (GAN) based methods (Goodfellow et al., 2020). Recently,

since the denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) was proposed for image-to-image translation, diffusion-based generative models have attracted much attention and surpassed GANs in many image generation tasks including image editing and image inpainting. A popular method called Stable-Diffusion was used to conduct a diffusion process in the low-dimensional latent space, and obtained high-resolution image synthesis results on various tasks such as super-resolution and inpainting, requiring relatively low computational costs (Rombach et al., 2022).

Because huge computational resources are required to adapt Transformer-based image and text encoders to other scenarios, the prompt-based visual fine-tuning method, named prompt tuning, is worth studying (Jia ML et al., 2022). It is well known that the Transformer-based vision encoders are significantly larger than their CNN counterparts, e.g., ViT-Huge (Dosovitskiy et al., 2021); therefore, we need an effective and efficient approach to better adapt these large models to downstream tasks rather than fully fine-tuning all the parameters. Inspired by prompting in NLP and VLMs, directly using the learnable prompts as Transformer input helps pre-trained ViTs achieve comparable performance with only 1% or fewer parameters to be tuned (Yao et al., 2021; Ge CJ et al., 2022; Ju et al., 2022).

As discussed above, VLMs have empowered various research fields, and we believe that they will further enable more diverse and new vision task formulations and model designs. Therefore, in this paper, we comprehensively review the recent advances in prompt learning for vision tasks after the explosion of VLMs and provide promising research directions. The overall structure of this article is shown in Fig. 2. In Section 2, we introduce some large-scale VLMs pre-trained with large numbers of image-text pairs and corresponding image-text contrastive learning schemes. In Section 3, we review the visual prompt learning methods, such as context optimization (CoOp), which enhanced classification performance against CLIP. In Section 4, we review recent popular prompt-guided generative models and relevant tasks such as image editing and inpainting. In Section 5, we show the efficient prompt/fine-tuning

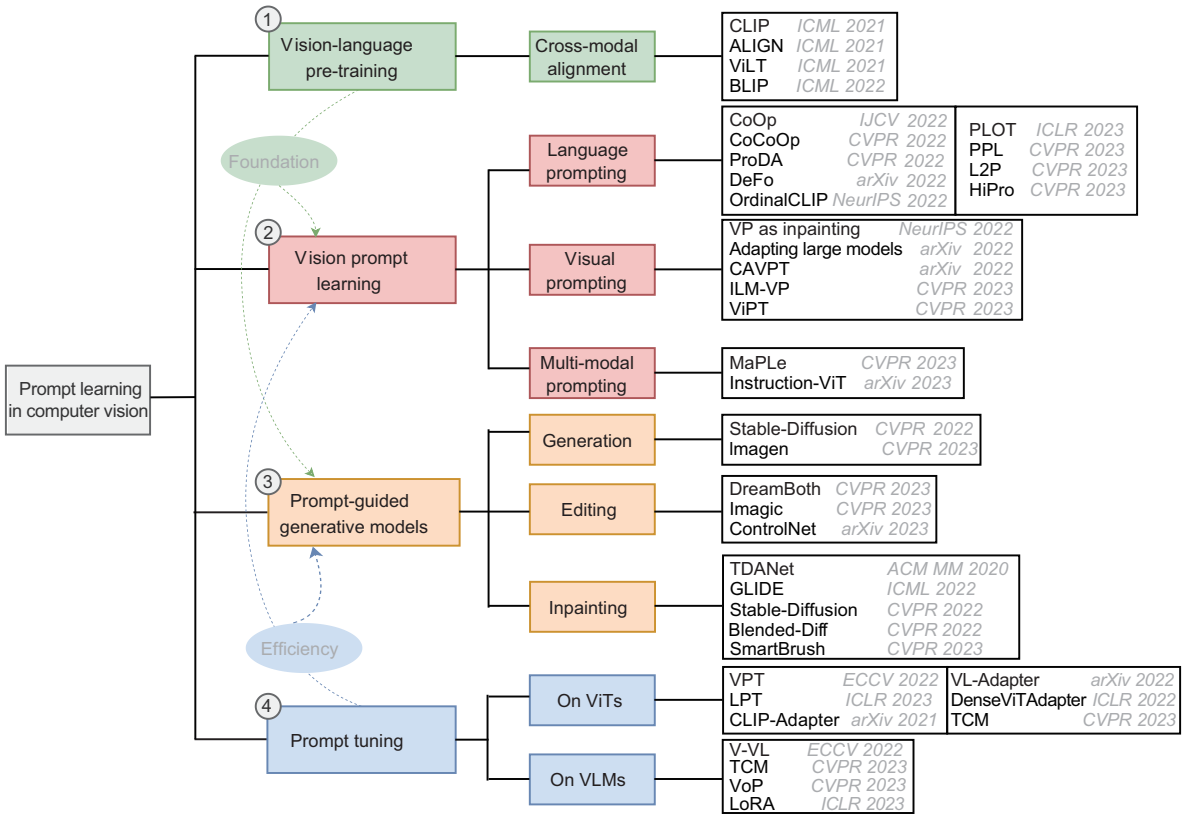

Fig. 2 Overall structure of this survey. We provide each research direction with some selected methods

strategies for adapting large ViTs to downstream tasks. In Section 6, we discuss some possible research directions integrated with prompt learning.

To collect the papers reviewed in this article, we entered the search terms "prompt learning," "visual prompting," "image generation/inpainting/editing with prompt," and "vision-language models" into the Web of Science with the time range from 2020 to 2023. With the survey, our motivation is to cover as many important papers as possible. Therefore, we searched for related articles using Google Scholar, PubMed, IEEE Xplore, and so on. Specifically, in Section 2, we choose mainly CLIP as the foundation model to introduce the current vision-language alignment paradigm. In Section 3, we choose the most popular basic methods of language prompting and visual prompting that have already been highly cited on Google Scholar, and then, we select application works that are correlated with these basic methods. In Section 4, we focus more on image-generation tasks and hotly discussed diffusion models such as Stable-Diffusion. In Section 5, we focus on prompt tuning on ViTs and VLMs, which is effective in improving the tuning efficiency of Transformer backbones and is critical to LLMs and VLMs. Consequently, the articles we selected are closely related to the prompt learning topic and highly related to artificial intelligence generated content (AIGC).

2 Vision-language pre-training

Current widely adopted AIGC techniques shine in innovatively and accurately injecting human language into generative models, making generated contents reflect meaningful semantic information. A critical process in such AIGC models is vision-language alignment that yields powerful and generative image and text encoders for various kinds of downstream tasks including discriminative and generative modeling. Therefore, before introducing visual prompt learning, we briefly review the fundamental VLMs and the corresponding contrastive learning scheme. As shown in Fig. 3, CLIP is the first large-scale pre-trained model for image-text alignment (Radford et al., 2021), and was pre-trained on four billion image-text pairs. ALIGN (Jia C et al., 2021) uses the same contrastive learning paradigm as used in CLIP to align images and text by training the Transformer encoders on a dataset even larger than that used in CLIP. The prediction probability of zero-shot CLIP is defined as follows:

$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, \boldsymbol{w}_y)/\tau)}{\sum_{i=1}^{K} \exp(\text{sim}(\boldsymbol{x}, \boldsymbol{w}_i)/\tau)}, \tag{1}$$

where $\boldsymbol{x}$ and $\{\boldsymbol{w}_i\}_{i=1}^{K}$ represent image features and text features generated by image and text encoders, respectively, K is the number of classes, $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity, and τ is the learnable temperature.

Bootstrapping language-image pre-training (BLIP) is a vision-language pre-training framework that transfers flexibly to both vision-language understanding and text generation. It effectively uses noisy web data by bootstrapping the captions, where a captioner can generate synthetic captions and a filter removes the noisy ones (Li JN et al., 2022).

Kim et al. (2021) proposed the vision-and-language Transformer (ViLT), which tackles two-modality information in a single unified manner.

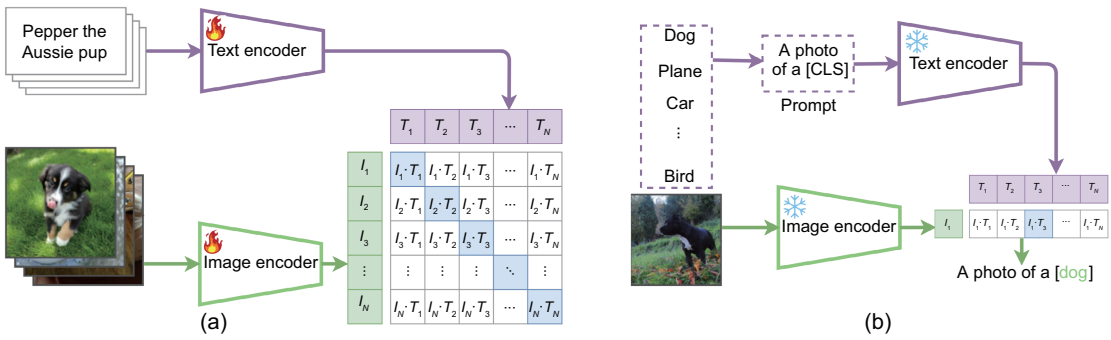

Fig. 3 Framework of contrastive language-image pre-training (CLIP): (a) contrastive pre-training using image-text pairs; (b) inference using the pre-trained image and text encoders (Reprinted from Radford et al. (2021), Copyright 2021, with permission from the authors)

It differs from previous VLM models mainly in its convolution-free embedding of pixel-level inputs. Removing deep text and vision embedders generates the lightweight visual inputs, which scales down the model size and speeds up the running. Simultaneously, it maintains comparable performance to PixelBert (Huang ZC et al., 2020) on downstream tasks.

Most of the existing VLMs share the same architecture of textual embedder-tokenizer from pre-trained BERT (Devlin et al., 2019). Such generalized text encoders with large numbers of parameters are trained on huge amounts of data, so prompt learning often fixes their pre-trained parameters, which will be discussed in the next section.

3 Vision prompt learning

Prompt learning has emerged as an effective technique for enhancing large pre-trained NLP models. Given its potential, researchers have become interested in studying learnable prompts in VLMs, such as CLIP, to enhance their generalizability and thereby achieve superior zero-shot and few-shot learning performance. In this section, we categorize current vision prompt learning methods into two categories, i.e., language prompting and visual prompting. In Table 1, we briefly summarize some selected works involving image-text alignment, language prompting, and visual prompting.

3.1 Language prompting

Language prompting refers to constructing learnable language contexts as text encoder inputs, and these tokens are trained during cross-modal contrastive learning. As summarized in Liu PF et al. (2023), prompt-based learning models the probability of text directly in language models, which is similar to that in vision models: the language prompting in vision models also aims at modeling text prompts as probabilities, and these can be discrete forms like "a photo of a [CLS]" or continuous forms of learnable vectors/variables (Zhou KY et al., 2022a). However, language prompting in vision tasks often depends on a language model to conduct vision-language alignment.

3.1.1 Methods and algorithms

CoOp was first proposed to introduce learnable prompt contexts, i.e., parameter vectors, to VLMs (Zhou KY et al., 2022b). The prediction probability is defined as follows:

$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_y))/\tau)}{\sum_{i=1}^{K} \exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_i))/\tau)}, \quad (2)$$

where $g(\cdot)$ denotes text encoder that is differentiable, $\boldsymbol{t}_i$ represents the prompt for the i^{th} class, i.e., $\boldsymbol{t}_i = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M, \boldsymbol{c}_i\}$, $\boldsymbol{v}_m$'s (m=1, 2, ..., M) are M learnable context vectors, and each of them has the same dimension as word embedding of $\boldsymbol{c}_i$.

Table 1 Summary and comparison of typical prompting methods

Method	Task	Language	Visual	Remark
CLIP (Radford et al., 2021)	Pre-train	✓	✗	Vision-language matching for pre-training
ALIGN (Jia C et al., 2021)	Pre-train	✓	✗	Vision-language matching with noisy data
BLIP (Li JN et al., 2022)	Pre-train	✓	✗	Removing noisy data with a bootstrapping captioner
CoOp (Zhou KY et al., 2022b)	Adapt.	✓	✗	Learnable prompts for adapting large-scale models
CoCoOp (Zhou KY et al., 2022a)	Adapt.	✓	✗	Conditional prompting for generalizing to new classes
DeFo (Wang F et al., 2023)	Adapt.	✓	✗	No class tokens in learnable prompts
PLOT (Chen GY et al., 2023)	Adapt.	✓	✗	Optimal transport
VP (Bahng et al., 2022)	Adapt.	✗	✓	Visual prompting with image perturbations
MAE-VQGAN (Bar et al., 2022)	Adapt.	✗	✓	Recovering visual prompt images via inpainting
Painter (Wang XL et al., 2023)	Seg./Det.	✗	✓	Proposing task prompts for many tasks
EVP (Liu WH et al., 2023)	Seg.	✗	✓	Task-aware explicit prompts for structure segmentation
ILM-VP (Chen AC et al., 2023)	Cls.	✗	✓	Visual prompting enhances traditional CNNs
BlackVIP (Oh et al., 2023)	Cls.	✗	✓	Predicting visual prompts for transfer learning
OrdinalCLIP (Li WH et al., 2022)	Ordinal reg.	✗	✓	Modeling ordinal regression as image-text matching
CLIP-Lung (Lei et al., 2023a)	Cls.	✗	✓	Channel-wise conditional prompts
MaPLe (Khattak et al., 2023)	Cls.	✓	✓	Predicting visual prompts for transfer learning
Instruction-ViT (Xiao et al., 2023)	Cls.	✓	✓	Tuning ViTs with multi-modal prompts
SAM (Kirillov et al., 2023)	Seg.	✓	✓	Prompts: points, boxes, texts

"Adapt.," "Seg.," "Det.," "Cls.," and "reg." are short for "Adaptation," "Segmentation," "Detection," "Classification," and "regression," respectively

CoOp fixes the CLIP pre-trained text encoder and updates the learnable contexts $\boldsymbol{v}_m$ without adding much training cost. Note that CoOp is a supervised learning framework in which gradients of context vectors are back-propagated from the final classifier using cross-entropy loss.

Although CoOp generalizes better than CLIP on downstream tasks, it suffers from performance degradation when encountering new classes; i.e., the CoOp learned prompts lack the ability to transfer to new classes. To address this weakness, Zhou KY et al. (2022a) proposed a conditional context optimization (CoCoOp), which enables the context learning process to be conditioned on visual features by introducing an additional meta-net for establishing relationships between latent visual features and learnable contexts. The prediction probability is defined as follows:

$$p(y|\boldsymbol{x})=\frac{\exp(\mathrm{sim}(\boldsymbol{x},g(\boldsymbol{t}_y(\boldsymbol{x})))/\tau)}{\sum_{i=1}^{K}\exp(\mathrm{sim}(\boldsymbol{x},g(\boldsymbol{t}_i(\boldsymbol{x})))/\tau)}, \quad (3)$$

where the context embeddings $\boldsymbol{t}_i$'s are conditioned on visual features, i.e., $\boldsymbol{t}_i=\{\boldsymbol{v}_1(\boldsymbol{x}), \boldsymbol{v}_2(\boldsymbol{x}),\dots,\boldsymbol{v}_M(\boldsymbol{x}),\boldsymbol{c}_i\}$, and $\boldsymbol{v}_m(\boldsymbol{x})=\boldsymbol{v}_m+\boldsymbol{\pi}$ (m=1, 2, …, M) where $\boldsymbol{\pi}=h_{\boldsymbol{\theta}}(\cdot)$ is the meta-net parameterized by $\boldsymbol{\theta}$. Different from CoOp, which updates only learnable prompt contexts, CoCoOp requires learning the additional parameters $\boldsymbol{\theta}$ of the proposed meta-net. Experimental results on the same 11 downstream tasks, such as Standford-Cars and FGVC-Aircraft, evaluated by CLIP and CoOp, have shown that CoCoOp exhibits improved performance on new classes over CoOp under the few-shot setting, confirming that instance-conditional prompts are capable of generalizing to new classes.

Despite the effectiveness of CoOp and CoCoOp in learning generalized prompt contexts, they are still constrained by class-specific prompts, and it is challenging for them to generate explainable context words. Different from CoOp and CoCoOp, decomposed feature prompting (DeFo) proposed by Wang F et al. (2023) abandons the prompt learning paradigm where the class tokens are contained in prompts. DeFo aims at decomposing visual features by applying learnable prompts. The key difference is that DeFo does not require the prompts contain specific class tokens, so each query image corresponds to n learnable query prompts, where n is a hyperparameter. However, DeFo needs an additional linear layer attached on top of the image encoder, which is used to map feature vectors to logits with the dimension of C, the number of classes. Note that the number of textual queries is independent of the number of classes in DeFo. During the training process, the textual queries and the additional linear layer are trainable for conducting vision-language alignment. Interestingly, the learned text embeddings of DeFo are more interpretable than those learned by CoOp and CoCoOp, where reasonable words can be obtained by retrieving a dictionary through similarity matching between learned prompts and vectors in this dictionary (Zhou KY et al., 2022b; Wang F et al., 2023).

Current prompt learning methods can prompt for the target image, but are less generalizable in the realistic scenario where one image contains multiple intrinsic attributes and extrinsic contexts that are difficult to fully describe with a single prompt. To address this issue, Chen GY et al. (2023) proposed to enhance prompt learning with optimal transport (PLOT) by aligning each image with multiple class-aware prompts. Specifically, PLOT holds the key distinction that the final prediction probability is obtained by measuring the optimal transport (OT) distance between visual features and multiple prompt vectors, which is defined as follows:

$$p(y|\boldsymbol{x})=\frac{\exp((1-d_{\mathrm{OT}}(y))/\tau)}{\sum_{k=1}^{K}\exp((1-d_{\mathrm{OT}}(k))/\tau)}, \quad (4)$$

where $d_{\mathrm{OT}}(k)$ is the OT distance obtained by inner-loop optimization using the Sinkhorn algorithm (Cuturi, 2013):

$$d_{\mathrm{OT}}(k)=d_{\mathrm{OT}}(\boldsymbol{u},\boldsymbol{v}|1-\boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}_k), \quad (5)$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are probability simplexes, and $\boldsymbol{F}$ and $\boldsymbol{G}$ are matrices of visual features and prompts, respectively. From Fig. 4, we can see that PLOT performs better on most of the downstream tasks than CoOp and CoCoOp. In addition, Selvaraju et al. (2017) showed that the learned multiple prompts indeed focus on varying regions in the images using Grad class activation mapping (CAM). Consequently, both quantitative and qualitative results demonstrate the effectiveness of aligning images with multiple prompts instead of a single prompt.

Prompt distribution learning, i.e., ProDA, was proposed to model prompt distribution for effectively adapting large VLMs to address downstream

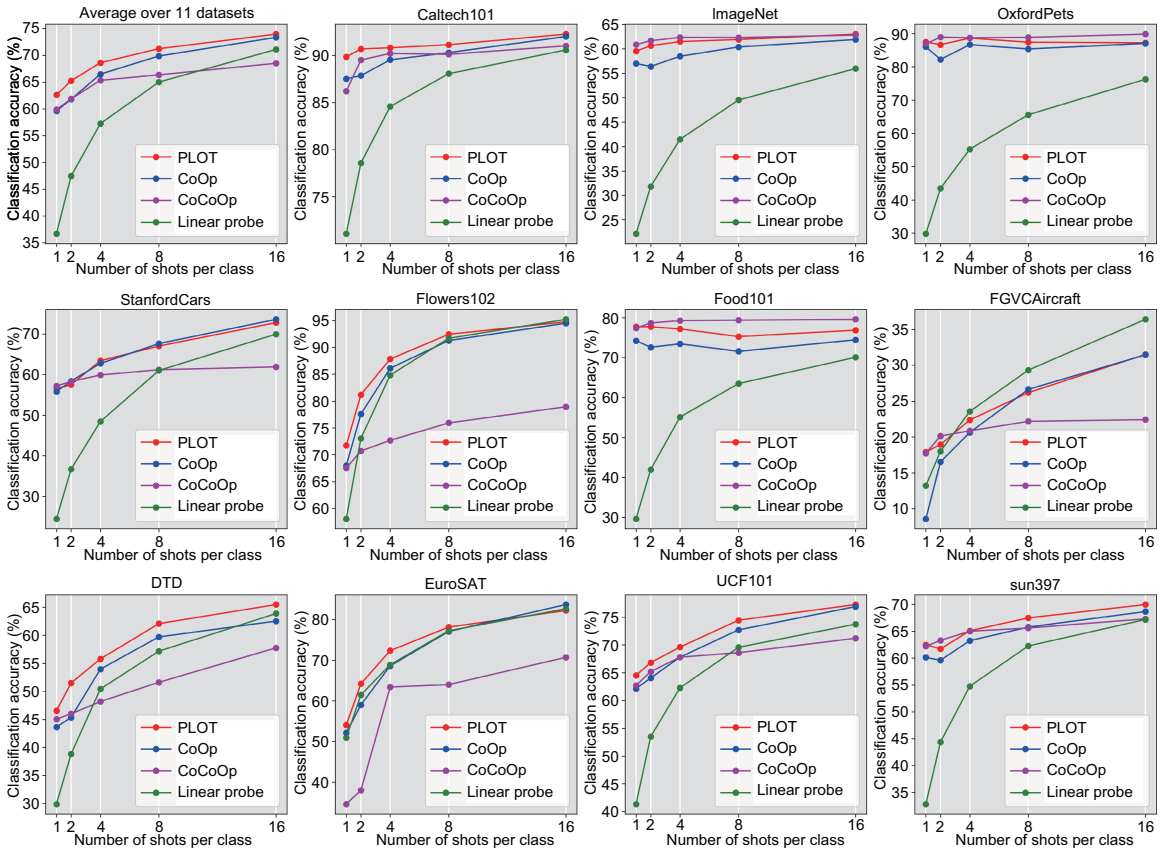

Fig. 4 Comparisons of classification accuracy on 11 downstream tasks (Reprinted from Chen GY et al. (2023), Copyright 2023, with permission from the authors)

tasks (Lu YN et al., 2022). ProDA learns low-bias prompts from a few samples and captures the distribution of diverse prompts simultaneously. This handles the varying visual representations during training. Probabilistic prompt learning (PPL) stems from the probability distribution perspective to address dense prediction tasks such as segmentation (Kwon et al., 2023); it aligns visual features and class-aware multiple prompts in the latent space via distribution alignment.

For continual learning, Wang ZF et al. (2022) proposed a novel approach, learning to prompt (L2P), to generate prompts that are effective in continual learning and can be used to adapt to new tasks without forgetting the previous ones. Hierarchical prompt (HiPro) learning was proposed for multi-task learning, which learns a set of hierarchical prompts that can be used to solve multiple tasks simultaneously. FedPR is a federated learning method for magnetic resonance imaging (MRI) reconstruction. It learns a set of visual prompts in the null space of the data distribution to improve the reconstruction performance. These three methods are all evaluated on various benchmarks and have shown promising results (Feng et al., 2023).

3.1.2 Applications of language prompting

Based on the prompt learning discussed above, the potential of this technique has been well explored. A successful application is modeling consistent ordinal information via vision-language alignment. Li WH et al. (2022) proposed OrdinalCLIP, which breaks the barrier of traditional ordinal regression methods that learn with categorical labels. OrdinalCLIP enforces visual features to approximate language prototypes of ranks via measuring Kullback-Leibler (KL) divergence between similarity matrices obtained by visual and language features. Although OrdinalCLIP does not achieve improved regression performance compared with state-of-the-art ordinal regression methods such as convolutional ordinal regression forest (CORF) for age

estimation (Zhu HP et al., 2022) and meta ordinal regression forest (MORF) for medical lesion progression (Lei et al., 2022), ordinal regression is a typical topic in machine learning, but it is still inferior in modeling non-equivalent ordinal relationships among visual features, which we argue is more realistic in real-world applications. Prompt learning has the potential to establish more real relationships among ranks via informative textual descriptions.

Recently, the segment anything model (SAM) has caught great attention in the CV community. Kirillov et al. (2023) built a huge dataset with 1.1 billion segmentation masks and trained the masked auto-encoder (MAE) based image encoder (He et al., 2022) and off-the-shelf CLIP-pre-trained text encoder in a supervised manner using dice loss (Milletari et al., 2016) and focal loss (Lin TY et al., 2017). SAM shows surprising capability in segmenting any unseen images. It allows users to provide multi-modal prompts like masks, points, boxes, and texts. This amazing model demonstrates the potential of prompts used in vision models and user-friendly application scenarios of multi-modal prompting.

CLIP-Lung was proposed using textual knowledge in medical datasets to facilitate medical image classification (Lei et al., 2023a). Specifically, it aligns textual attributes and class texts of lung nodules to the corresponding image features, and then generates fine-grained attention maps correlated with small lesions. Yu Y et al. (2022) applied prompt learning for COVID-19 diagnosis through modeling the conditional probability of vocabulary conditioned on learnable prompts. Language prompting also enhances interpretability of deep neural networks through global and local alignments between language and image features (Lei et al., 2023b).

In summary, language prompting provides visual features and image encoders with extra learning targets. We argue that the trainable language contexts act as noisy class information, informally, a Gaussian distribution centered at the specific classes, which helps improve the generalizability and robustness of visual features.

3.2 Visual prompting

Visual prompting refers to constructing learnable image tokens, i.e., visual perturbations as shown in Fig. 5, as input of an image encoder, and these learnable tokens are trained to adapt large-scale vision models. As we can see, this kind of learnable visual prompt plays a role similar to language prompts in LLMs, where the prompted model will perform a new task with perturbations.

3.2.1 Methods and algorithms

Visual prompting is used in the field of CV and adapts pre-trained visual models to new downstream tasks without requiring task-specific fine-tuning or modifications to the model architecture (Lin Y et al., 2023). It was first defined in Bahng et al. (2022) to mimic the prompting idea in NLP. Fig. 5 illustrates the framework of visual prompting. It has been explored under different names, such as model reprogramming or adversarial reprogramming in the CV domain. These techniques aim to repurpose pre-trained vision models for new tasks by leveraging universal input patterns and output label mappings. Visual prompting offers the advantage of parameter-efficient fine-tuning, requiring less parameter storage and training data compared to full fine-tuning. It has been applied to various scenarios, including black-box source models, cross-domain transfer learning, and improving metrics like adversarial robustness and fairness.

One approach to visual prompting treats the problem as image inpainting. The goal is to fill in the missing parts of a concatenated visual prompt image, which is created by combining input-output image examples. Bar et al. (2022) demonstrated that leveraging an appropriate inpainting algorithm trained on a curated dataset can lead to surprisingly effective results. They trained a masked auto-encoder on an unlabeled dataset comprising figures from academic papers and applied visual prompting to pre-trained models for various image-to-image tasks such as foreground segmentation, single object detection, colorization, and edge detection.

To investigate the efficacy of visual prompting in adapting large-scale models in CV, built upon recent approaches like prompt tuning and adversarial reprogramming, Bahng et al. (2022) proposed to learn a single image perturbation that enables a frozen model to perform a new task. The authors demonstrated that visual prompting, particularly for models like CLIP, is effective and robust to distribution shift, achieving competitive performance over standard linear probes. They also analyzed various aspects such as the downstream dataset, prompt

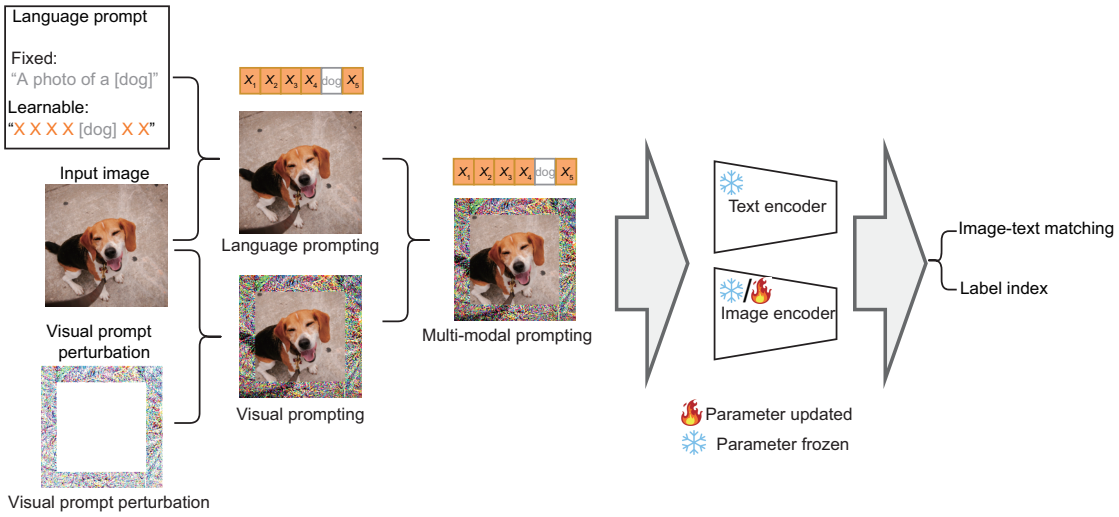

Fig. 5 Hierarchical relationships among different prompting methods: language prompting, visual prompting with image perturbations, and multi-modal prompting

design, and output transformation in relation to adaptation performance.

Many approaches have attempted to prompt with images, similar to prefix tuning in NLP. For instance, colorful prompt tuning (CPT) focuses on extending the capabilities of language-based models by training vision encoders and using continuous image embeddings as visual prefixes or colored blocks and color-based textual prompts for visual grounding (Yao et al., 2021). However, these approaches primarily aim to enhance language models rather than specifically investigating the efficacy of visual prompting in enhancing vision models.

To address the adaptation of pre-trained VLMs like CLIP to downstream tasks via prompt tuning, Xing et al. (2022) proposed a novel dual-modality prompt tuning (DPT) paradigm that simultaneously learns text and visual prompts. They introduced a class-aware visual prompt tuning (CAVPT) scheme, which dynamically generates class-aware visual prompts by performing cross-attention between text prompt features and image patch token embeddings. This encoding captures both task-related and visual instance information, ultimately improving the concentration of the final image feature on the target visual concept. Experimental results on 11 datasets demonstrated the effectiveness and generalization ability of the proposed method.

Chen AC et al. (2023) delved into the technique of visual prompting and explored its effectiveness given ruleless label mapping between source and target classes. The authors proposed the iterative label mapping based visual prompting (ILM-VP) framework, which automatically re-maps source labels to target labels to improve the accuracy of visual prompting on target tasks. Additionally, they integrated label mapping into the text prompt selection of CLIP to enhance its target task accuracy.

Visual prompting and model reprogramming have gained attention in CV research. Visual prompt tuning (VPT), a related technique restricted to ViTs, enables to learn prompting parameters at intermediate layers of a source model and to extract modality-related prompts, which has enhanced multi-modal tracking tasks. Visual prompt multi-modal tracking (ViPT) was proposed to learn a few modal-relevant prompts for adapting large ViTs to downstream tracking tasks (Zhu JW et al., 2023).

3.2.2 Applications of visual prompting

Painter is based on the idea of task prompts (Wang XL et al., 2023). Taking semantic segmentation as an example, task prompts indicate the concatenations of input images and the corresponding segmentation target. Then the target inputs are also concatenations of images and blank images to be predicted. Furthermore, Painter can tackle various tasks such as object detection, deraining, and depth estimation. Liu WH et al. (2023) proposed explicit visual prompting to simultaneously address low-level structure segmentations, including

camouflaged object, forgery, shadow, and defocus blur detection. To enhance the performance of traditional CNN models on target tasks like Flowers102, Chen AC et al. (2023) proposed to couple the design of label mapping with visual prompt training, which significantly improved the classification accuracy of CNNs such as ResNet. To build a consistent relationship between language and visual information in image classification, Mao et al. (2023) proposed a new object recognition benchmark that requires the model to simultaneously predict the right categorical labels and the rationales; this rationale knowledge is transferred from LLMs such as GPT-3. For transfer learning, Oh et al. (2023) proposed black-box visual prompting (BlackVIP) to adapt pre-trained visual models to downstream tasks, which beats traditional fine-tuning and visual fine-tuning strategies.

Hence, visual prompting is complementary to language prompting for enhancing image encoders, and is also regarded as perturbation or augmentation of input images with respect to class texts. Furthermore, the combination of these two modalities of prompts enables the interaction between vision and language information while improving performance in unseen classes or tasks.

3.3 Multi-modal prompting

Based on the success of language and visual prompt learning, the combinations of these two modalities of prompts have caught much attention in recent years. MaPLe is a novel prompt learning framework that uses both vision and language modules of CLIP to improve alignment between these two representations. It demonstrates state-of-the-art results toward unseen new categories and domain transfer. The design promotes strong coupling between the cross-modal prompts to ensure mutual synergy and discourages learning independent unimodal solutions (Khattak et al., 2023). Instruction-ViT conducts multi-modal prompts for instruction learning. It uses prompt tuning to adapt pre-trained models to different downstream tasks, reducing the number of trainable parameters while improving the performance on unknown tasks (Xiao et al., 2023). Unified-IO is a framework that integrates various vision tasks, including image classification, object detection, depth estimation, and image inpainting (Lu JS et al., 2023). We can see that appropriate cooperation among various prompt modalities can facilitate a wide range of applications.

Consequently, prompts have been proven to play a crucial role in LLMs, and in recent years, vision models have also been using prompts to improve scalability for multiple downstream tasks.

4 Prompt-guided generative models

Deep generative models are extensively trained to approximate complicated, high-dimensional probability distributions. The trained models can then be used to estimate the likelihood of observations and to create new samples from the underlying distribution. Popular models include GAN (Goodfellow et al., 2020) and variational auto-encoder (VAE) (Kingma and Welling, 2013). Recently, diffusion-based generative models (Ho et al., 2020) have attracted much attention and outperform traditional GAN-based methods in both quantitative and qualitative image generation results. Furthermore, because large VLMs can effectively incorporate linguistic prompts into vision models, prompt-guided image generation theories and applications have been developed and garnered considerable attention from both the academic and industrial communities.

Conventional text-condition GAN related methods have been widely studied (Reed et al., 2016a, 2016b; Zhang H et al., 2017; Xu T et al., 2018; Tao et al., 2022), without interacting with humans or language information. The advanced prompt-based generation models have emerged as a prominent approach for enhancing human-in-the-loop conditional generation in vision-language models. Notable applications such as Stable-Diffusion (Rombach et al., 2022), DALL-E (Ramesh et al., 2021, 2022), and MidJourney (https://en.wikipedia.org/wiki/Midjourney) have been developed and achieved widespread recognition globally.

Different from previous techniques, these models allow users to employ high-level semantic information, including natural language, stick figures, and binary masks, as prompts to generate desired outcomes flexibly. Owing to their versatility and user-friendliness, prompt-guided generative models have emerged as a hotly discussed research direction, encompassing various fields such as prompt-guided image generation and editing, image inpainting, and three-dimensional object generation.

The success of prompt-guided generative models can be attributed to three aspects:

1. The pre-trained large-scale multi-modal models. The large-scale VLMs, such as CLIP (Radford et al., 2021), BLIP (Li JN et al., 2022), and ViLT (Kim et al., 2021), and speech-language models including Whisper (Radford et al., 2023), provide researchers with powerful image and text encoders. These models capture high-level contextual information effectively, facilitating efficient connections between multi-modal prompts and images. Consequently, the optimization process in the training phase is streamlined as the model can focus on conditional generation or editing tasks rather than learning to extract context.

2. The advanced diffusion-based generative models. In contrast to conventional GAN models that often specialize in specific topics such as human faces (Abdal et al., 2019; Karras et al., 2019, 2020, 2021), diffusion models offer enhanced flexibility and strong unconditional generation ability. Such features enable the generation of diverse scenes and facilitate large-scale training, avoiding the limitations of topic-specific approaches.

3. The cross-attention mechanism which has demonstrated its power in harnessing contextual information from different modalities (Lin HZ et al., 2022). Being advantageous because it is easily implemented, cross-attention enables the integration of prompts from diverse modalities, thus significantly enhancing the performance and capabilities of prompt-guided image generation models.

Next, we summarize and discuss image generation, image editing, and image inpainting tasks. Some popular methods are summarized in Table 2.

4.1 Image generation

A highly successful approach for prompt-guided image generation is the latent diffusion model (Rombach et al., 2022), also known as Stable-Diffusion, which performs diffusion in the latent space of pre-trained auto-encoders and generates images iteratively to reduce inference time. During training, the prompts of various modalities are first encoded by a domain-specific network and then interact with the intermediate layers of the diffusion model by cross-attention mechanisms (Vaswani et al., 2017; Lin HZ et al., 2022).

Fig. 6 shows the architectures of two popular latent diffusion models, Stable-Diffusion (Rombach et al., 2022) and Imagen (Wang S et al., 2023), which share similar designs. Both models perform diffusion in the latent space, resulting in a considerable reduction of computational costs. Stable-Diffusion supports several kinds of prompts, including semantic maps, text, representations, and images. For each modality, a domain-specific encoder is used to translate the prompt for the generation model; such an encoder could be either a pre-trained model or the one that is to be trained simultaneously with the diffusion model. GLIDE is a text-guided diffusion model that enables text-conditional image synthesis for photorealistic image generation (Nichol et al., 2022). The authors explored two kinds of guidance, CLIP guidance and classifier-free guidance.

Although diffusion models have emerged as new state-of-the-art large-scale generative models, recent works have made attempts to train GAN with a similar strategy by scaling up the model and leveraging text prompts. Kang et al. (2023) introduced GigaGAN, a novel GAN architecture that avoids previous limits and showcases GANs as a feasible choice for text-to-image synthesis. GigaGAN achieves competitive performance against diffusion models while significantly reducing inference time, making it an efficient option for prompt-to-image synthesis.

These fundamental generation models have been employed to accomplish a diverse array of astounding downstream tasks, including image editing, image inpainting, and others.

4.2 Image editing

Image editing refers to the process of altering an image to make it more visually appealing or to convey a specific message. Generally, image editing involves removing unwanted regions like specks and scratches, rotating and cropping images, correcting for lens aberrations, sharpening or softening the image, making color changes, and adding special effects to the image.

Traditional image editing methods are almost all based on GANs (Perarnau et al., 2016; Wang TC et al., 2018; Ling et al., 2021). For text-guided diffusion models, DreamBooth allows users to personalize text-to-image models by fine-tuning them with just a few images of a subject (Ruiz et al., 2023). Given just a few images of target subjects as input, DreamBooth fine-tunes a pre-trained text-to-image model to bind a unique identifier, i.e., a unique prompt with a specific subject. ControlNet presents a framework

Table 2 Summary and comparison of typical generative models

Method	Task	Language	Visual	Remark
Stable-Diffusion (Rombach et al., 2022)	Generation	✓	✓	Diffusion-based, many kinds of prompts
Imagen (Wang S et al., 2023)	Generation	✓	✗	Diffusion-based
GLIDE (Nichol et al., 2022)	Generation	✓	✗	Text-conditional image synthesis
GigaGAN (Kang et al., 2023)	Generation	✓	✗	A new GAN-based model trained with prompts
DreamBooth (Ruiz et al., 2023)	Editing	✓	✗	Unique prompt for a specific subject
ControlNet (Zhang LM et al., 2023)	Editing	✓	✗	Task-specific conditions, robust to small datasets
Blended-Diff (Avrahami et al., 2022)	Inpainting	✓	✗	Task-specific conditions, robust to small datasets
SmartBrush (Xie et al., 2023)	Inpainting	✓	✗	Continuous guidance, multi-level masks

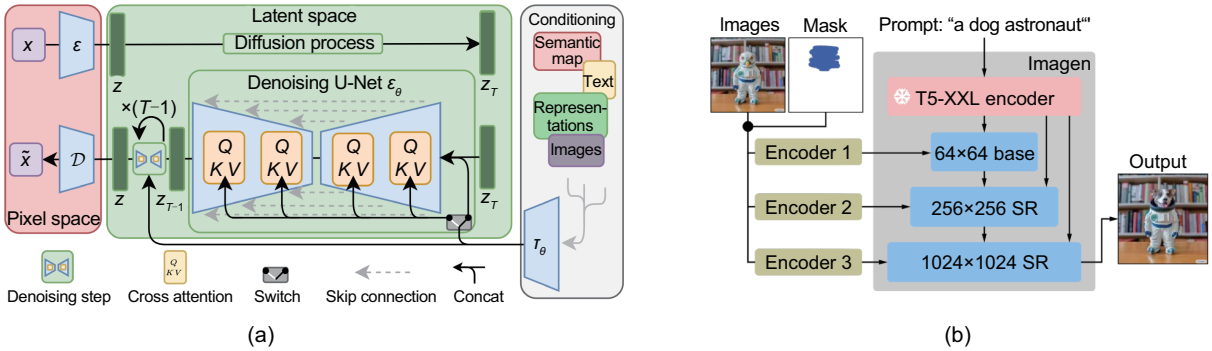

Fig. 6 Prompt-guided image generation framework: (a) Stable-Diffusion (Reprinted from Rombach et al. (2022), Copyright 2022, with permission from IEEE); (b) Imagen (Reprinted from Wang S et al. (2023), Copyright 2023, with permission from IEEE). These two models use a similar design, incorporating the cross-attention mechanism to fuse the prompt into the generation model (i.e., the latent diffusion model in both cases)

to control pre-trained large diffusion models to support additional input conditions. ControlNet learns task-specific conditions, and the learning is robust to small datasets. Adding conditional control to text-to-image diffusion models allows users to guide image generation more precisely, resulting in more accurate and tailored outputs (Zhang LM et al., 2023). The Imagic method is also a text-conditioned diffusion model; it differs from previous methods in requiring only one image of one object for training (Kawar et al., 2023).

In summary, diffusion models with prompt conditions satisfy the nature of language-intervened image editing where the models can generate mimic results relative to target prompt inputs.

4.3 Image inpainting

Image inpainting is a technology that aims to fill in the missing regions of an image based on the available surrounding pixels while conforming to human visual perception and cognition. With the advancements in deep learning techniques, image inpainting has emerged as a prominent research field. In contrast to traditional approaches that rely on similarity-based filling using neighboring pixels or patches (Barnes et al., 2009; Xu ZB and Sun, 2010; He and Sun, 2014; Lee et al., 2016), deep learning based methods have the ability to perform dynamic restoration at the pixel level. These methods often employ a two-stage framework, wherein the image is restored in a coarse-to-fine manner, addressing the limitation of a restricted receptive field (Iizuka et al., 2017; Yu JH et al., 2018, 2019). To address the coherent limitation of the receptive field, self-attention models and Fourier-based convolutional networks have been introduced (Zhou YQ et al., 2021; Suvorov et al., 2022). By incorporating larger or global receptive fields, these network architectures effectively capture the undisturbed structural information presented in images. Although the current methods have achieved excellent results, these architectures lack control of the results, particularly when dealing with large holes where existing areas may not provide sufficient contextual information. Recent advancements in multi-modal inpainting methods have shown significant improvements by allowing models to inpaint using prompts such as class labels, text descriptions, and masks.

One approach for prompt-guided inpainting is TDANet, which trains networks end to end and involves aligning image-text pairs (Zhang ZJ et al., 2020). However, this kind of method often yields suboptimal results in aligning high-level contexts with low-level details. Building on the recent success of diffusion-based conditional and text-to-image generation models, researchers have proposed to directly extend the unconditional DDPM model to image inpainting tasks by performing resampling at each reverse step, thereby transforming the task into unconditional generation (Lugmayr et al., 2022). Inspired by its remarkable generalizability, blended diffusion (Blended-Diff) (Avrahami et al., 2022) adapts the pre-trained unconditional diffusion model and employs the CLIP score to encourage the output to align with the provided text prompts, as illustrated in Fig. 7b. As shown in Fig. 8, Imagen successfully edites the puppy using both mask and language prompts. Blended-Diff effectively integrates language priors from CLIP to generate the inpainted object. Smartbooth further emphasizes the importance of the mask prompt to constrain the shape of the inpainted object.

On the other hand, further fine-tuning the model can potentially lead to high-quality results, albeit at the expense of increased computational cost. Approaches such as GLIDE (Nichol et al., 2022), Stable-Diffusion (Rombach et al., 2022), and SmartBrush (Xie et al., 2023) employ supervised fine-tuning of text-to-image conditional diffusion models for image inpainting tasks, specifically focusing on inpainting the correct contextual information within the masked regions. SmartBrush (Xie et al., 2023) introduces a novel approach that enables continuous guidance of the inpainting process using text and object shapes, allowing for different levels of mask precision, as shown in Fig. 7a. This allows users to inpaint images with provided masks, such as the specific shape of a cat. The distinctive aspect of SmartBrush lies in its training process, which involves using instance segmentation masks and local text descriptions from a pre-trained BLIP (Li JN et al., 2022). As a result, SmartBrush not only facilitates precise image inpainting but also leverages the mask precision information from a trained U-Net. This allows SmartBrush to effectively preserve the background even without a precise mask during the inference stage. The qualitative and quantitative evaluations demonstrate that SmartBrush achieves superior results in terms of visual quality, mask controllability, and background preservation, as presented in Fig. 8c.

Hence, a notable advantage of prompt-guided inpainting using diffusion models is its ability to accurately recover missing regions by leveraging prior prompts, which is a challenge to overcome with conventional inpainting methods.

5 Prompt tuning

Although prompt learning has achieved amazing progress recently, there exists an inevitable problem, i.e., the difficulty for general users to train such large models from scratch due to the requirement of huge computation resources. Therefore, researchers began to explore new parameter-efficient tuning methods for better using its representation ability. In this section, we will introduce prompt tuning methods that have great potential for general users to benefit from pre-trained knowledge. Table 3 illustrates the summarization and comparison of typical prompt tuning methods.

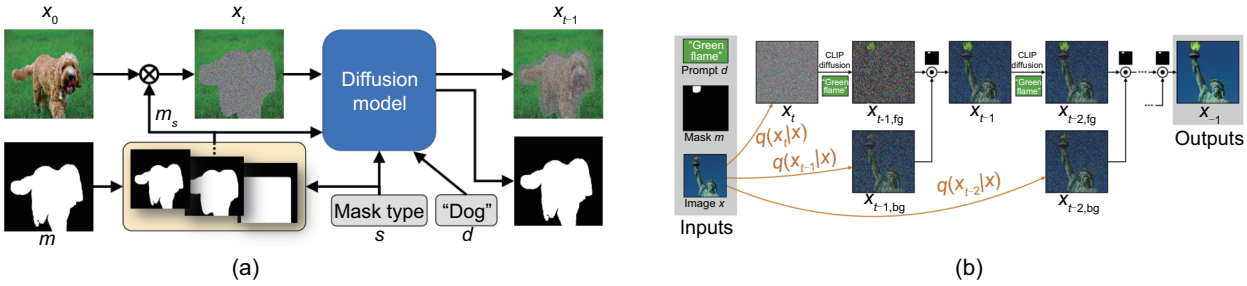

Fig. 7 Two popular prompt-guided inpainting frameworks: (a) SmartBrush (Reprinted from Xie et al. (2023), Copyright 2023, with permission from IEEE), which uses language and mask as prompts; (b) Blended-Diff (Reprinted from Avrahami et al. (2022), Copyright 2022, with permission from IEEE), which uses language as a prompt on a pre-trained DDPM model

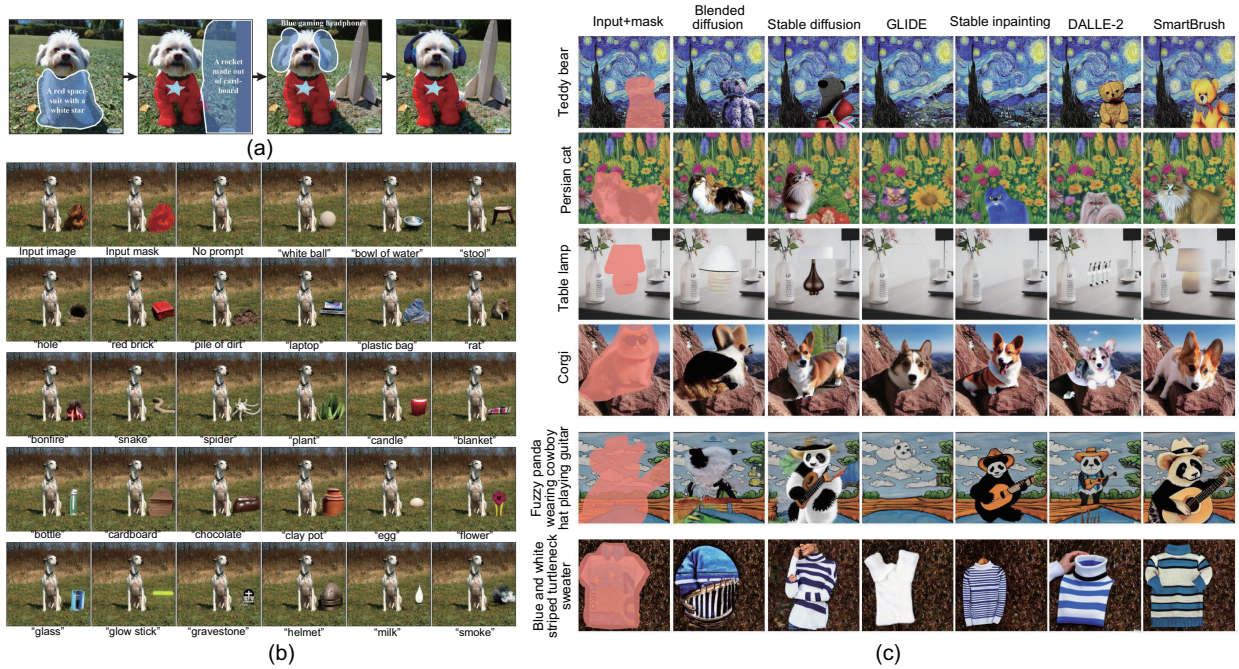

Fig. 8 Image inpainting results: (a) Imagen (Reprinted from Wang S et al. (2023), Copyright 2023, with permission from IEEE); (b) Blended-Diff (Reprinted from Avrahami et al. (2022), Copyright 2022, with permission from IEEE); (c) SmartBrush (Reprinted from Xie et al. (2023), Copyright 2023, with permission from IEEE). The prompt-based inpainting method significantly improves interactivity and enables easy image editing through prompts

Table 3 Summary and comparison of typical prompt tuning methods

Method	Task	Language	Visual	Remark
VPT (Jia ML et al., 2022)	ViTs	✓	✗	Less than 1% of backbone parameters to tune
LPT (Dong et al., 2023)	ViTs	✓	✗	Prompt tuning for long-tailed classification
TCM (Yu WW et al., 2023)	VLMs	✓	✗	Scene text detection with two prompt generators
V-VL (Ju et al., 2022)	VLMs	✓	✗	Video understanding using learnable prompts
VoP (Huang ST et al., 2023)	VLMs	✓	✗	Video understanding, prompts in intermediate layers
DenseVitAdapter (Chen Z et al., 2023)	VLMs	✓	✗	Injecting task- and input-specific knowledge
CLIP-Adapter (Gao et al., 2021)	VLMs	✓	✗	Modality-agnostic adapter
VL-Adapter (Sung et al., 2022)	VLMs	✓	✗	Unifying various image-text and video-text downstream tasks
LoRA (Hu et al., 2022)	VLMs	✓	✗	Parameter-efficient training on downstream tasks

5.1 Prompt tuning on ViTs

Scaling laws (Kaplan et al., 2020) drive models to larger scales. While larger ViT models typically perform better, their generalization ability remains limited. This means that even big models need to be fine-tuned on downstream tasks to achieve satisfactory performance. However, fully fine-tuning the entire pre-trained visual model on different tasks and data distributions can be computationally expensive. Therefore, it is imperative to study parameter-efficient tuning methods for pre-trained models, such as ViT, to reduce computational costs and improve efficiency.

VPT was proposed to efficiently fine-tune ViTs by posing trainable prompt vectors at the input layer of ViTs (Jia ML et al., 2022). The authors proposed two versions: VPT-Deep, which includes trainable prompts in the intermediate layers of ViTs, and VPT-Shallow, which does not. Note that the total number of fine-tuned VPT parameters is smaller than 1% of the number of backbone parameters. For the classification performances on downstream tasks, VPT surpasses adaptor-based strategies (Houlsby et al., 2019; Pfeiffer et al., 2020a, 2020b).

VPT for generative transfer learning was proposed to learn a set of visual prompts that can be used to improve the transferability of generative

models. The method is evaluated on various benchmarks and has shown promising results (Sohn et al., 2023).

Learning with long-tailed data is a challenging problem in machine learning/deep learning, where the imbalanced data distribution of different classes dramatically guides the model to overfit major classes while being less generative to tailed classes. Existing works fine-tune the models pre-trained on large-scale unlabeled datasets, suffering from high costs in computation and overfitting to long-tailed data. To address this issue, Dong et al. (2023) proposed long-tailed prompt tuning (LPT) to adapt large VLMs to long-tailed image classification tasks. LPT is a two-stage framework in which shared prompts at the shared prompt tuning stage, patch tokens, and class tokens are fine-tuned, while the group prompt tuning stage is used for learning group-specific prompts by matching class-token-grouped prompts. LPT surpasses VPT in long-tailed tasks using different backbone networks including ResNet-152 (He et al., 2016) and ViT-B (Dosovitskiy et al., 2021).

5.2 Prompt tuning on VLMs

VLMs leverage semantically rich verbal information to learn visual information under language supervision. Models trained in this manner exhibit strong transferability and achieve performance improvements, particularly in zero-shot and few-shot scenarios. Prompt tuning can also be employed to efficiently fine-tune VLMs.

TCM (Yu WW et al., 2023) is a novel approach to scene text detection that leverages prompt learning to enable flexible model adaptation. The method uses two elaborate prompt generators that plug into a pre-trained VLM, allowing the image and text embeddings to interact and fill in missing information in single modality embeddings. To further improve the quality of the embeddings, an additional matching loss is proposed to constrain the similarity between the curated embeddings. The experiments demonstrate that TCM has great transferability, making it a promising approach for scene text detection.

V-VL (Ju et al., 2022) is a prompt-based learning method that guides a pre-trained image-based VLM toward video understanding tasks, such as action recognition, action localization, and text-video retrieval. To achieve this goal, learnable prompts are appended to the text-tokenized results and then fed to the text encoder. Furthermore, V-VL employs a Transformer encoder based temporal modeling method to aggregate the image features. This approach enables the frozen pre-trained image-based VLM to adapt well to various video understanding tasks. The experiments conducted on zero-shot and few-shot scenarios demonstrate strong generalization.

VoP focuses on adapting pre-trained CLIP to video-text retrieval tasks in a parameter-efficient manner. Huang ST et al. (2023) proposed visual-text dual prompts and inserted them in each layer of frozen encoders. To address spatial-temporal modeling without introducing additional modules, they proposed three different visual prompts conditioned on position, context, and function, separately. This approach enables the pre-trained model on images to be transferred to the video domain. With only 0.1% trainable prompts, VoP achieves a 1.4% average relative improvement compared to full fine-tuning. Overall, VoP provides an efficient and effective way to adapt pre-trained models to video-understanding tasks.

As we can see, prompt tuning has shown great potential for transferring pre-trained models to a wide range of downstream tasks. Adapters (Houlsby et al., 2019) can be used as an alternative method to achieve parameter-efficient tuning. DenseVit-Adapter (Chen Z et al., 2023) focuses on injecting both task- and input-specific knowledge into the pre-trained model through a lightweight adapter. CLIP-Adapter (Gao et al., 2021) is a modality-agnostic adapter that allows VLMs to balance pre-training learned knowledge with newly learned knowledge from few-shot samples. VL-Adapter (Sung et al., 2022) unifies various image- and video-text downstream tasks in a single vision and language framework, enabling efficient fine-tuning of the pre-trained VLM with adapters in a multi-task setting. In addition to these methods, LoRA (Hu et al., 2022) can be trained on downstream tasks in a parameter-efficient manner and has shown great potential in the field of CV (Smith et al., 2023).

Hence, prompt tuning is an effective tool for landing specific downstream tasks by adapting learned knowledge of large models.

6 Future directions

Although visual prompt learning has been widely studied and applied in many applications, it remains in its early stages of development. This section reviews and highlights some promising directions in prompt learning.

6.1 Image classification

Based on the paradigm of classification via image-text matching brought from cross-modal contrastive learning, such as CLIP, there has been a notable enhancement in the performance of image classification benchmarks, especially in zero- and few-shot settings. In addition, prompt learning has shown critical effects in adapting pre-trained models to downstream tasks and specific scenarios, like generalization to new classes (Zhou KY et al., 2022a). However, to further unlock the potential of prompt learning for image classification, two primary challenges persist: (1) Adapting or training large-scale models to specific scenarios suffers from domain shifts between target texts and pre-training textual data; (2) More explainable learned prompts are required to facilitate the interpretation of classification models, which is the key to explainable AI in the future. Hence, prompt learning will be effective in real applications and in interpreting AI models.

6.2 Semantic segmentation

The current popular segmentation model may be SAM (Kirillov et al., 2023), which is a large model trained on huge numbers of image-prompt pairs. In addition to its amazing performance, it allows multiple kinds of prompts for the users to choose, including points, boxes, and texts. Recently, more and more studies have shown that SAM fails in some real scenarios, like medical image segmentation. This poses a valuable research topic on integrating domain-specific knowledge into SAM-like large segmentation models. We think that efficient prompt learning, which can connect language and target regions in the images, will enable a human-in-the-loop segmentation paradigm to sense human knowledge.

6.3 Open-vocabulary object detection

Open-vocabulary object detection (OVOD) differs from traditional object detection in that the detectors of OVOD are trained on base classes existing in the training set, and should be able to recognize new classes. Vision and language knowledge distillation (ViLD) directly employs the CLIP text encoder to obtain text embedding, which is subsequently used to classify object proposals and act as supervision of training detectors (Gu et al., 2022). Because ViLD is a two-stage detection method, Ma et al. (2022) proposed HierKD, a one-stage OVOD detection framework, which distills the knowledge of CLIP at global and instance levels. For novel classes, HierKL outperms previous you only look once (YOLO) based zero-shot detectors. To avoid the obstacle of handcrafted prompts used in ViLD, Du et al. (2022) incorporated learnable prompts into OVOD, and the proposed DetPro significantly improved performances upon ViLD under various experimental settings. Although OVOD has been explored toward applying knowledge of LLMs, we still need an elegant way of reducing distribution shift between the base and new classes, even for more specific scenarios such as medicine, remote sensing, and auto-driving, where the text annotations are difficult to acquire.

6.4 Multi-task learning

Multi-task learning is a classical problem that aims at performing multiple tasks using a single model or a single ensemble model, i.e., multiple outputs corresponding to various tasks such as classification, regression, and reconstruction. Based on multi-modal prompting, multi-task learning intrinsically correlates well with constructing multi-modal prompts; then we need to develop effective learning targets and efficient training strategies to enhance the performance of multiple tasks. Another kind of multi-task learning refers to multiple distributions of training data; for example, Liu YJ et al. (2023) proposed HiPro learning, which targets at solving a new benchmark that involves different kinds of classification datasets including fine-grained recognition like Flowers102, scene recognition, action recognition, and general recognition. This work triggers the building of fine-grained relationships between tasks and corresponding prompts in a real application; e.g., multiple diseases in a clinic should possess their own specific prompts.

6.5 Chain-of-thought

Chain-of-thought (CoT) is a concept in LLMs, where researchers found that eliciting complex multi-step reasoning through step-by-step answer examples can significantly improve the ability of LLMs to perform complex reasoning, as shown in Fig. 9 (Kojima et al., 2022; Wei et al., 2022; Zhang ZS et al., 2022). CoT prompting allows LLMs to "infer the answer like a human" instead of generating answers directly. There are two primary paradigms in CoT prompting: one employs a straightforward prompt to facilitate step-by-step reasoning in a zero-shot manner, and the other presents several manual demonstrations sequentially in a few-shot approach. Ge JX et al. (2023) applied CoT in prompt tuning, which differed from those prompt learning methods like CoOp and CoCoOp, because they made a chain of prompts before the text encoder, with all the prompts in the CoT chain being conditioned on visual features, and this generalized better in new classes than CoCoOp. Because language information is the key to CoT, multi-modal prompting will be a more promising way of using the CoT concept. For example, researchers have explored CoT for multi-modal reasoning in science question answering (Lu P et al., 2022).

In conclusion, CoT is a powerful and innovative concept that enhances the performance and capabilities of large-scale models, leading to improved context awareness and semantic consistency by enabling better understanding and connection of diverse ideas and concepts. On the other hand, with the rapid development of VLMs and LLMs, CoT is an effective tool for enhancing vision models by injecting CoT into a wide range of applications of language- or rationale-specific scenarios like medicine and auto-driving.

6.6 Prompt for medical image analysis

Medical data are intrinsically involved with images of different modalities and corresponding text or language annotations. Low-level medical image processing such as denoising and deblurring acts as a critical part of high-level tasks such as classification and segmentation. It is worthwhile to explore how class-related textual information can affect joint learning of high- and low-level tasks, and whether the prompt embeddings are beneficial for the performances of both tasks especially for disease diagnosis (Lei et al., 2021; Xu ZH et al., 2022).

On the other hand, multi-modal prompting can effectively leverage language and visual information, which will help the model learn consistent feature spaces. Therefore, this is helpful for explainable AI for medical diagnoses.

6.7 Prompt for weather forecasting

In weather forecasting tasks, the accumulated massive NLP-formatted data in the past could not be effectively used. It is a valuable research direction to combine prompt learning to enable forecasting models to better understand weather conditions

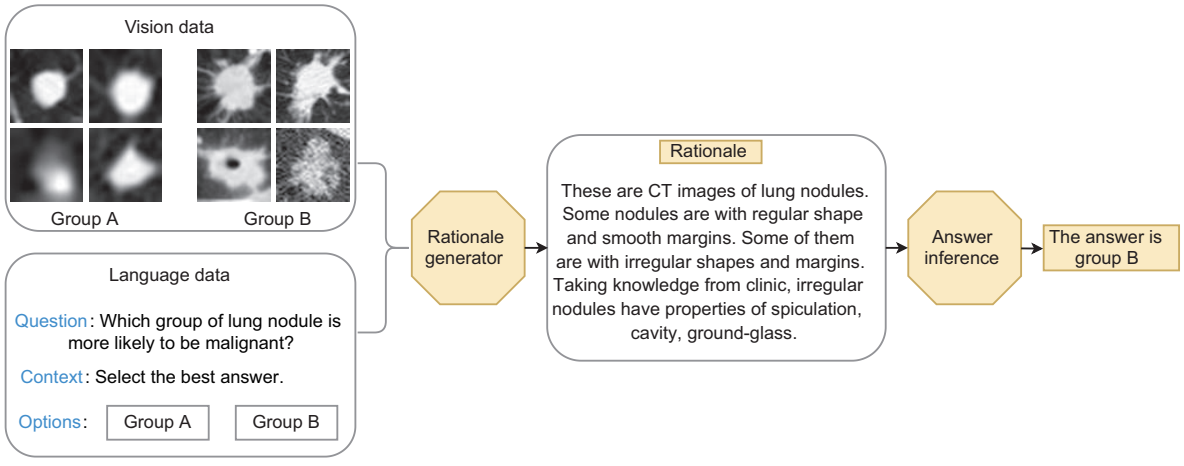

Fig. 9 The framework of multi-modal chain-of-thought (CoT) (Zhang ZS et al., 2023), an example of lung nodule malignancy recognition using computed tomography (CT) images. The framework of CoT can be divided into two stages: rationale generation and answer inference

by leveraging textual knowledge.

Furthermore, the results of time-series prediction tasks are often limited, typically consisting of only future numerical values. By incorporating prompt learning in conjunction with NLP, it becomes possible to introduce a wider variety of supervisory information. Xue and Salim (2022) proposed prompt-based time-series forecasting (PromptCast) and achieved superior prediction performance compared to various neural network models in tasks such as temperature prediction. In the context of precipitation nowcasting, Cao et al. (2023) demonstrated that adjusting the learning order and weights of multiple forecasting tasks can significantly improve the prediction performance of nowcasting models, highlighting the importance of task formulation. The integration of prompt learning with meteorological forecasting is a highly promising research direction.

6.8 Prompt for gait recognition

Gait recognition involves using the biological gait pattern to identify an individual's identity. One of the main challenges in this field is extracting robust and discriminative gait features (Lin BB et al., 2021; Chao et al., 2022). A very recent work, MaskCL (Li MK et al., 2023), introduces gait silhouettes as semantic prompts for the person re-identification task. This allows the model to learn some cross-clothes invariant features. In the gait recognition task, researchers have studied cross-view robust gait features (Lin BB et al., 2021; Li JQ et al., 2023a, 2023b) for better recognition performance, which is a significant challenge in this field. Designing gait prompts to learn cross-view robust features is a promising direction for future research in this area.

7 Conclusions

In this survey, we have comprehensively reviewed studies with respect to prompt learning in the CV field. We first introduced the prompt learning strategies in image-text cross-modal frameworks, which have overcome shortcomings like weak generalizability of traditional prompt engineering. Then, for efficiently applying large VLMs for downstream prompt-guided learning, we discussed prompt tuning methods tailored for better adaptation of large-scale ViTs. Through richer information contained in prompts, prompt-guided generative models trigger a wide variety of applications such as image generation, image editing, and image inpainting. Finally, we provided some perspectives on prompt learning in the CV field and possible and promising research directions.

In summary, prompt learning has great potential in enhancing existing studies and leading to new multi-modal directions.

Contributors

Yiming LEI and Hongming SHAN designed the structure and logic of the paper. Yiming LEI drafted the whole paper. Yuan CAO reviewed the visual prompt learning part. Zilong LI reviewed the prompt-guided generative models part. Jingqi LI reviewed the prompt tuning part. Yiming LEI and Hongming SHAN revised and finalized the paper. All the authors proofread the paper.

Compliance with ethics guidelines

All the authors declare that they have no conflict of interest.

References

Abdal R, Qin YP, Wonka P, 2019. Image2StyleGAN: how to embed images into the StyleGAN latent space? Proc IEEE/CVF Int Conf on Computer Vision, p.4431-4440. https://doi.org/10.1109/ICCV.2019.00453

Avrahami O, Lischinski D, Fried O, 2022. Blended diffusion for text-driven editing of natural images. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18187-18197. https://doi.org/10.1109/CVPR52688.2022.01767

Bahng H, Jahanian A, Sankaranarayanan S, et al., 2022. Exploring visual prompts for adapting large-scale models. https://doi.org/10.48550/arXiv.2203.17274

Bar A, Gandelsman Y, Darrell T, et al., 2022. Visual prompting via image inpainting. Proc 36th Conf on Neural Information Processing Systems, p.25005-25017.

Barnes C, Shechtman E, Finkelstein A, et al., 2009. PatchMatch: a randomized correspondence algorithm for structural image editing. *ACM Trans Graph*, 28(3):24. https://doi.org/10.1145/1531326.1531330

Cao Y, Zhang DC, Zheng X, et al., 2023. Mutual information boosted precipitation nowcasting from radar images. *Remote Sens*, 15(6):1639. https://doi.org/10.3390/rs15061639

Chao HQ, Wang K, He YW, et al., 2022. GaitSet: cross-view gait recognition through utilizing gait as a deep set. *IEEE Trans Patt Anal Mach Intell*, 44(7):3467-3478. https://doi.org/10.1109/TPAMI.2021.3057879

Chen AC, Yao YG, Chen PY, et al., 2023. Understanding and improving visual prompting: a label-mapping perspective. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.19133-19143. https://doi.org/10.1109/CVPR52729.2023.01834

Chen GY, Yao WR, Song XC, et al., 2023. PLOT: prompt learning with optimal transport for vision-language models. Proc 11th Int Conf on Learning Representations.

Chen Z, Duan YC, Wang WH, et al., 2023. Vision Transformer adapter for dense predictions. Proc 11th Int Conf on Learning Representations.

Cuturi M, 2013. Sinkhorn distances: lightspeed computation of optimal transport. Proc 26th Int Conf on Neural Information Processing Systems, p.2292-2300. https://doi.org/10.5555/2999792.2999868

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional Transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.4171-4186. https://doi.org/10.18653/v1/N19-1423

Dong BW, Zhou P, Yan SC, et al., 2023. LPT: long-tailed prompt tuning for image classification. Proc 11th Int Conf on Learning Representations.

Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2021. An image is worth 16×16 words: Transformers for image recognition at scale. Proc 9th Int Conf on Learning Representations.

Du Y, Wei FY, Zhang ZH, et al., 2022. Learning to prompt for open-vocabulary object detection with vision-language model. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.14064-14073. https://doi.org/10.1109/CVPR52688.2022.01369

Feng CM, Li BJ, Xu XX, et al., 2023. Learning federated visual prompt in null space for MRI reconstruction. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8064-8073. https://doi.org/10.1109/CVPR52729.2023.00779

Gao P, Geng SJ, Zhang RR, et al., 2021. CLIP-Adapter: better vision-language models with feature adapters. https://doi.org/10.48550/arXiv.2110.04544

Ge CJ, Huang R, Xie MX, et al., 2022. Domain adaptation via prompt learning. https://doi.org/10.48550/arXiv.2202.06687

Ge JX, Luo HY, Qian SY, et al., 2023. Chain of thought prompt tuning in vision language models. https://doi.org/10.48550/arXiv.2304.07919

Goodfellow I, Pouget-Abadie J, Mirza M, et al., 2020. Generative adversarial networks. *Commun ACM*, 63(11):139-144. https://doi.org/10.1145/3422622

Gu XY, Lin TY, Kuo WC, et al., 2022. Open-vocabulary object detection via vision and language knowledge distillation. Proc 10th Int Conf on Learning Representations.

Han K, Wang YH, Chen HT, et al., 2023. A survey on vision Transformer. *IEEE Trans Patt Anal Mach Intell*, 45(1):87-110. https://doi.org/10.1109/TPAMI.2022.3152247

He KM, Sun J, 2014. Image completion approaches using the statistics of similar patches. *IEEE Trans Patt Anal Mach Intell*, 36(12):2423-2435. https://doi.org/10.1109/TPAMI.2014.2330611

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

He KM, Chen XL, Xie SN, et al., 2022. Masked autoencoders are scalable vision learners. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15979-15988. https://doi.org/10.1109/CVPR52688.2022.01553

Ho J, Jain A, Abbeel P, 2020. Denoising diffusion probabilistic models. Proc 34th Int Conf on Neural Information Processing Systems, p.574. https://doi.org/10.5555/3495724.3496298

Houlsby N, Giurgiu A, Jastrzebski S, et al., 2019. Parameter-efficient transfer learning for NLP. Proc 36th Int Conf on Machine Learning, p.2790-2799.

Hu EJ, Shen YL, Wallis P, et al., 2022. LoRA: low-rank adaptation of large language models. Proc 10th Int Conf on Learning Representations.

Huang ST, Gong B, Pan YL, et al., 2023. VoP: text-video co-operative prompt tuning for cross-modal retrieval. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6565-6574. https://doi.org/10.1109/CVPR52729.2023.00635

Huang ZC, Zeng ZY, Liu B, et al., 2020. Pixel-BERT: aligning image pixels with text by deep multi-modal Transformers. https://doi.org/10.48550/arXiv.2004.00849

Iizuka S, Simo-Serra E, Ishikawa H, 2017. Globally and locally consistent image completion. *ACM Trans Graph*, 36(4):107. https://doi.org/10.1145/3072959.3073659

Jia C, Yang YF, Xia Y, et al., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. Proc 38th Int Conf on Machine Learning, p.4904-4916.

Jia ML, Tang LM, Chen BC, et al., 2022. Visual prompt tuning. Proc 17th European Conf on Computer Vision, p.709-727. https://doi.org/10.1007/978-3-031-19827-4_41

Ju C, Han TD, Zheng KH, et al., 2022. Prompting visual-language models for efficient video understanding. Proc 17th European Conf on Computer Vision, p.105-124. https://doi.org/10.1007/978-3-031-19833-5_7

Kang M, Zhu JY, Zhang R, et al., 2023. Scaling up GANs for text-to-image synthesis. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10124-10134. https://doi.org/10.1109/CVPR52729.2023.00976

Kaplan J, McCandlish S, Henighan T, et al., 2020. Scaling laws for neural language models. https://doi.org/10.48550/arXiv.2001.08361

Karras T, Laine S, Aila T, 2019. A style-based generator architecture for generative adversarial networks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4396-4405. https://doi.org/10.1109/CVPR.2019.00453

Karras T, Laine S, Aittala M, et al., 2020. Analyzing and improving the image quality of StyleGAN. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8107-8116. https://doi.org/10.1109/CVPR42600.2020.00813

Karras T, Aittala M, Laine S, et al., 2021. Alias-free generative adversarial networks. Proc 35th Conf on Neural Information Processing Systems, p.852-863.

Kawar B, Zada S, Lang O, et al., 2023. Imagic: text-based real image editing with diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6007-6017. https://doi.org/10.1109/CVPR52729.2023.00582

Khan S, Naseer M, Hayat M, et al., 2022. Transformers in vision: a survey. *ACM Comput Surv*, 54(10s):200. https://doi.org/10.1145/3505244

Khattak MU, Rasheed H, Maaz M, et al., 2023. MaPLe: multi-modal prompt learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.19113-19122. https://doi.org/10.1109/CVPR52729.2023.01832

Kim W, Son B, Kim I, 2021. ViLT: vision-and-language Transformer without convolution or region supervision. Proc 38th Int Conf on Machine Learning, p.5583-5594.

Kingma DP, Welling M, 2013. Auto-encoding variational Bayes. https://doi.org/10.48550/arXiv.1312.6114

Kirillov A, Mintun E, Ravi N, et al., 2023. Segment anything. https://doi.org/10.48550/arXiv.2304.02643

Kojima T, Gu SS, Reid M, et al., 2022. Large language models are zero-shot reasoners. Proc 36th Conf on Neural Information Processing Systems.

Kwon H, Song T, Jeong S, et al., 2023. Probabilistic prompt learning for dense prediction. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6768-6777. https://doi.org/10.1109/CVPR52729.2023.00654

Lee JH, Choi I, Kim MH, 2016. Laplacian patch-based image synthesis. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2727-2735. https://doi.org/10.1109/CVPR.2016.298

Lei YM, Zhang JP, Shan HM, 2021. Strided self-supervised low-dose CT denoising for lung nodule classification. *Phenomics*, 1(6):257-268. https://doi.org/10.1007/s43657-021-00025-y

Lei YM, Zhu HP, Zhang JP, et al., 2022. Meta ordinal regression forest for medical image classification with ordinal labels. *IEEE/CAA J Autom Sin*, 9(7):1233-1247. https://doi.org/10.1109/JAS.2022.105668

Lei YM, Li ZL, Shen Y, et al., 2023a. CLIP-Lung: textual knowledge-guided lung nodule malignancy prediction. Proc 26th Int Conf on Medical Image Computing and Computer-Assisted Intervention, p.403-412. https://doi.org/10.1007/978-3-031-43990-2_38

Lei YM, Li ZL, Li YY, et al., 2023b. LICO: explainable models with language-image consistency. https://doi.org/10.48550/arXiv.2310.09821

Li JN, Li DX, Xiong CM, et al., 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. Proc 39th Int Conf on Machine Learning, p.12888-12900.

Li JQ, Gao JQ, Zhang YZ, et al., 2023a. Motion matters: a novel motion modeling for cross-view gait feature learning. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5. https://doi.org/10.1109/ICASSP49357.2023.10096571

Li JQ, Zhang YZ, Shan HM, et al., 2023b. Gaitcotr: improved spatial-temporal representation for gait recognition with a hybrid convolution-Transformer framework. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5. https://doi.org/10.1109/ICASSP49357.2023.10096602

Li MK, Xu P, Li CG, et al., 2023. MaskCL: semantic mask-driven contrastive learning for unsupervised person re-identification with clothes change. https://doi.org/10.48550/arXiv.2305.13600

Li WH, Huang XK, Zhu Z, et al., 2022. OrdinalCLIP: learning rank prompts for language-guided ordinal regression. Proc 36th Conf on Neural Information Processing Systems.

Lin BB, Zhang SL, Yu X, 2021. Gait recognition via effective global-local feature representation and local temporal aggregation. Proc IEEE/CVF Int Conf on Computer Vision, p.14628-14636. https://doi.org/10.1109/ICCV48922.2021.01438

Lin HZ, Cheng X, Wu XY, et al., 2022. CAT: cross attention in vision Transformer. Proc IEEE Int Conf on Multimedia and Expo, p.1-6. https://doi.org/10.1109/ICME52920.2022.9859720

Lin TY, Goyal P, Girshick R, et al., 2017. Focal loss for dense object detection. Proc IEEE Int Conf on Computer Vision, p.2999-3007. https://doi.org/10.1109/ICCV.2017.324

Lin Y, Zhao ZC, Zhu ZJ, et al., 2023. Exploring visual prompts for whole slide image classification with multiple instance learning. https://doi.org/10.48550/arXiv.2303.13122

Ling H, Kreis K, Li DQ, et al., 2021. EditGAN: high-precision semantic image editing. Proc 35th Conf on Neural Information Processing Systems, p.16331-16345.

Liu PF, Yuan WZ, Fu JL, et al., 2023. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv*, 55(9):195. https://doi.org/10.1145/3560815

Liu WH, Shen X, Pun CM, et al., 2023. Explicit visual prompting for low-level structure segmentations. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.19434-19445. https://doi.org/10.1109/CVPR52729.2023.01862

Liu YJ, Lu YN, Liu H, et al., 2023. Hierarchical prompt learning for multi-task learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10888-10898. https://doi.org/10.1109/CVPR52729.2023.01048

Lu JS, Clark C, Zellers R, et al., 2023. Unified-IO: a unified model for vision, language, and multi-modal tasks. Proc 11th Int Conf on Learning Representations.

Lu P, Mishra S, Xia T, et al., 2022. Learn to explain: multimodal reasoning via thought chains for science question answering. Proc 36th Conf on Neural Information Processing Systems, p.2507-2521.

Lu YN, Liu JZ, Zhang YG, et al., 2022. Prompt distribution learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5196-5205. https://doi.org/10.1109/CVPR52688.2022.00514

Lugmayr A, Danelljan M, Romero A, et al., 2022. Repaint: inpainting using denoising diffusion probabilistic models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11451-11461. https://doi.org/10.1109/CVPR52688.2022.01117

Ma ZY, Luo G, Gao J, et al., 2022. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.14054-14063. https://doi.org/10.1109/CVPR52688.2022.01368

Mao CZ, Teotia R, Sundar A, et al., 2023. Doubly right object recognition: a why prompt for visual rationales. Proc IEEE/CVF Conf on Computer Vision and Pattern

Recognition, p.2722-2732.
https://doi.org/10.1109/CVPR52729.2023.00267

Milletari F, Navab N, Ahmadi SA, 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation. Proc 4th Int Conf on 3D Vision, p.565-571. https://doi.org/10.1109/3DV.2016.79

Nichol AQ, Dhariwal P, Ramesh A, et al., 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. Proc 39th Int Conf on Machine Learning, p.16784-16804.

Oh C, Hwang H, Lee HY, et al., 2023. BlackVIP: black-box visual prompting for robust transfer learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.24224-24235.
https://doi.org/10.1109/CVPR52729.2023.02320

Perarnau G, van de Weijer J, Raducanu B, et al., 2016. Invertible conditional GANs for image editing.
https://doi.org/10.48550/arXiv.1611.06355

Pfeiffer J, Kamath A, Rücklé A, et al., 2020a. AdapterFusion: non-destructive task composition for transfer learning. Proc 16th Conf of the European Chapter of the Association for Computational Linguistics: Main Volume, p.487-503.
https://doi.org/10.18653/v1/2021.eacl-main.39

Pfeiffer J, Rücklé A, Poth C, et al., 2020b. AdapterHub: a framework for adapting Transformers. Proc Conf on Empirical Methods in Natural Language Processing: System Demonstrations, p.46-54.
https://doi.org/10.18653/v1/2020.emnlp-demos.7

Radford A, Kim JW, Hallacy C, et al., 2021. Learning transferable visual models from natural language supervision. Proc 38th Int Conf on Machine Learning, p.8748-8763.

Radford A, Kim JW, Xu T, et al., 2023. Robust speech recognition via large-scale weak supervision. Proc 40th Int Conf on Machine Learning, p.28492-28518.

Ramesh A, Pavlov M, Goh G, et al., 2021. Zero-shot text-to-image generation. Proc 38th Int Conf on Machine Learning, p.8821-8831.

Ramesh A, Dhariwal P, Nichol A, et al., 2022. Hierarchical text-conditional image generation with CLIP latents.
https://doi.org/10.48550/arXiv.2204.06125

Reed S, Akata Z, Yan XC, et al., 2016a. Generative adversarial text to image synthesis. Proc 33rd Int Conf on Machine Learning, p.1060-1069.
https://doi.org/10.5555/3045390.3045503

Reed S, Akata Z, Mohan S, et al., 2016b. Learning what and where to draw. Proc 30th Int Conf on Neural Information Processing Systems, p.217-225.
https://doi.org/10.5555/3157096.3157121

Rombach R, Blattmann A, Lorenz D, et al., 2022. High-resolution image synthesis with latent diffusion models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10674-10685.
https://doi.org/10.1109/CVPR52688.2022.01042

Ruiz N, Li YZ, Jampani V, et al., 2023. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22500-22510.
https://doi.org/10.1109/CVPR52729.2023.02155

Selvaraju RR, Cogswell M, Das A, et al., 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proc IEEE Int Conf on Computer Vision, p.618-626.
https://doi.org/10.1109/ICCV.2017.74

Shamshad F, Khan S, Zamir SW, et al., 2023. Transformers in medical imaging: a survey. *Med Image Anal*, 88:102802.
https://doi.org/10.1016/j.media.2023.102802

Smith JS, Hsu YC, Zhang LY, et al., 2023. Continual diffusion: continual customization of text-to-image diffusion with C-LoRA.
https://doi.org/10.48550/arXiv.2304.06027

Sohn K, Chang HW, Lezama J, et al., 2023. Visual prompt tuning for generative transfer learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.19840-19851.
https://doi.org/10.1109/CVPR52729.2023.01900

Sung YL, Cho J, Bansal M, 2022. VL-Adapter: parameter-efficient transfer learning for vision-and-language tasks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5217-5227.
https://doi.org/10.1109/CVPR52688.2022.00516

Suvorov R, Logacheva E, Mashikhin A, et al., 2022. Resolution-robust large mask inpainting with Fourier convolutions. Proc IEEE/CVF Winter Conf on Applications of Computer Vision, p.3172-3182.
https://doi.org/10.1109/WACV51458.2022.00323

Tao M, Tang H, Wu F, et al., 2022. DF-GAN: a simple and effective baseline for text-to-image synthesis. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16494-16504.
https://doi.org/10.1109/CVPR52688.2022.01602

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.6000-6010.
https://doi.org/10.5555/3295222.3295349

Wang F, Li ML, Lin XD, et al., 2023. Learning to decompose visual features with latent textual prompts. Proc 11th Int Conf on Learning Representations.

Wang S, Saharia C, Montgomery C, et al., 2023. Imagen Editor and EditBench: advancing and evaluating text-guided image inpainting. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.18359-18369.
https://doi.org/10.1109/CVPR52729.2023.01761

Wang TC, Liu MY, Zhu JY, et al., 2018. High-resolution image synthesis and semantic manipulation with conditional GANs. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8798-8807.
https://doi.org/10.1109/CVPR.2018.00917

Wang XL, Wang W, Cao Y, et al., 2023. Images speak in images: a generalist painter for in-context visual learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6830-6839.
https://doi.org/10.1109/CVPR52729.2023.00660

Wang ZF, Zhang ZZ, Lee CY, et al., 2022. Learning to prompt for continual learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.139-149. https://doi.org/10.1109/CVPR52688.2022.00024

Wei J, Wang XZ, Schuurmans D, et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Proc 36th Conf on Neural Information Processing Systems.

Xiao ZX, Chen YZ, Zhang L, et al., 2023. Instruction-ViT: multi-modal prompts for instruction learning in ViT. https://doi.org/10.48550/arXiv.2305.00201

Xie SA, Zhang ZF, Lin Z, et al., 2023. SmartBrush: text and shape guided object inpainting with diffusion model. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22428-22437. https://doi.org/10.1109/CVPR52729.2023.02148

Xing YH, Wu QR, Cheng D, et al., 2022. Class-aware visual prompt tuning for vision-language pre-trained model. https://doi.org/10.48550/arXiv.2208.08340

Xu T, Zhang PC, Huang QY, et al., 2018. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1316-1324. https://doi.org/10.1109/CVPR.2018.00143

Xu ZB, Sun J, 2010. Image inpainting by patch propagation using patch sparsity. *IEEE Trans Image Process*, 19(5):1153-1165. https://doi.org/10.1109/TIP.2010.2042098

Xu ZH, Shen B, Tang YL, et al., 2022. Deep clinical phenotyping of Parkinson's disease: towards a new era of research and clinical care. *Phenomics*, 2(5):349-361. https://doi.org/10.1007/s43657-022-00051-4

Xue H, Salim FD, 2022. Prompt-based time series forecasting: a new task and dataset. http://export.arxiv.org/abs/2210.08964v1

Yao Y, Zhang A, Zhang ZY, et al., 2021. CPT: colorful prompt tuning for pre-trained vision-language models. https://doi.org/10.48550/arXiv.2109.11797

Yu JH, Lin Z, Yang JM, et al., 2018. Generative image inpainting with contextual attention. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5505-5514. https://doi.org/10.1109/CVPR.2018.00577

Yu JH, Lin Z, Yang JM, et al., 2019. Free-form image inpainting with gated convolution. Proc IEEE/CVF Int Conf on Computer Vision, p.4470-4479. https://doi.org/10.1109/ICCV.2019.00457

Yu WW, Liu YL, Hua W, et al., 2023. Turning a CLIP model into a scene text detector. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6978-6988. https://doi.org/10.1109/CVPR52729.2023.00674

Yu Y, Rong L, Wang MY, et al., 2022. Prompt learning for multi-modal COVID-19 diagnosis. Proc IEEE Int Conf on Bioinformatics and Biomedicine, p.2803-2807. https://doi.org/10.1109/BIBM55620.2022.9995157

Zhang H, Xu T, Li HS, et al., 2017. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. Proc IEEE Int Conf on Computer Vision, p.5908-5916. https://doi.org/10.1109/ICCV.2017.629

Zhang LM, Rao A, Agrawala M, 2023. Adding conditional control to text-to-image diffusion models. https://doi.org/10.48550/arXiv.2302.05543

Zhang ZJ, Zhao Z, Zhang Z, et al., 2020. Text-guided image inpainting. Proc 28th ACM Int Conf on Multimedia, p.4079-4087. https://doi.org/10.1145/3394171.3413939

Zhang ZS, Zhang A, Li M, et al., 2022. Automatic chain of thought prompting in large language models. Proc 11th Int Conf on Learning Representations.

Zhang ZS, Zhang A, Li M, et al., 2023. Multimodal chain-of-thought reasoning in language models. https://doi.org/10.48550/arXiv.2302.00923

Zhou KY, Yang JK, Loy CC, et al., 2022a. Conditional prompt learning for vision-language models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16795-16804. https://doi.org/10.1109/CVPR52688.2022.01631

Zhou KY, Yang JK, Loy CC, et al., 2022b. Learning to prompt for vision-language models. *Int J Comput Vis*, 130(9):2337-2348. https://doi.org/10.1007/s11263-022-01653-1

Zhou YQ, Barnes C, Shechtman E, et al., 2021. TransFill: reference-guided image inpainting by merging multiple color and spatial transformations. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2266-2267. https://doi.org/10.1109/CVPR46437.2021.00230

Zhu HP, Shan HM, Zhang YH, et al., 2022. Convolutional ordinal regression forest for image ordinal estimation. *IEEE Trans Neur Netw Learn Syst*, 33(8):4084-4095. https://doi.org/10.1109/TNNLS.2021.3055816

Zhu JW, Lai SM, Chen X, et al., 2023. Visual prompt multi-modal tracking. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9516-9526. https://doi.org/10.1109/CVPR52729.2023.00918