Prompt-based Adaptation in Large-scale Vision Models: A Survey

Anonymous authors
Paper under double-blind review

Abstract

In computer vision, Visual Prompting (VP) and Visual Prompt Tuning (VPT) have recently emerged as lightweight and effective alternatives to full fine-tuning for adapting large-scale vision models within the "pretrain-then-finetune" paradigm. However, despite rapid progress, their conceptual boundaries remain blurred, as VP and VPT are frequently used interchangeably in current research, reflecting a lack of systematic distinction between these techniques and their respective applications. In this survey, we revisit the designs of VP and VPT from first principles, and conceptualize them within a unified framework termed Prompt-based Adaptation (PA). Within this framework, we distinguish methods based on their injection granularity: VP operates at the pixel level, while VPT injects prompts at the token level. We further categorize these methods by their generation mechanism into fixed, learnable, and generated prompts. Beyond the core methodologies, we examine PA's integrations across diverse domains, including medical imaging, 3D point clouds, and vision-language tasks, as well as its role in test-time adaptation and trustworthy AI. We also summarize current benchmarks and identify key challenges and future directions. To the best of our knowledge, we are the first comprehensive survey dedicated to PA's methodologies and applications in light of their distinct characteristics. Our survey aims to provide a clear roadmap for researchers and practitioners in all area to understand and explore the evolving landscape of PA-related research.

1 Introduction

Large-scale vision models, exemplified by the Vision Transformer (ViT) (Dosovitskiy et al., 2021) and Swin Transformer (Liu et al., 2021), have fundamentally transformed computer vision. These models are typically pretrained on massive datasets (*e.g.*, ImageNet-21k (Russakovsky et al., 2015)) to acquire transferable representations, which can subsequently be finetuned for specific downstream tasks (Iofinova et al., 2022) (*e.g.*, FGVC (Jia et al., 2022), VTAB-1k (Zhai et al., 2019)). This approach is commonly referred to as the "pretrain-then-finetune" paradigm, which can markedly reduce the reliance on labeled data (Han et al., 2024). As the scale of these models continues to grow (Han et al., 2023), conventional full fine-tuning (FT), which updates all parameters, has become increasingly costly in terms of computation and storage, and risks eroding valuable pretrained knowledge Han et al. (2024). In response, a variety of parameter-efficient fine-tuning (PEFT) methods, aiming to finetune models by adjusting only a small fraction of parameters while keeping the remainder frozen, have been developed. Among these, Prompt-based Adaptation (PA) has emerged as a particularly prominent and effective technique (Jia et al., 2022).

In this survey, we provide a systematic review and categorization of recent PA algorithms and their practical implementations. Unlike existing surveys, which primarily focus on multimodal or vision–language settings, our work centers exclusively on PA within vision models. Understanding the confusing definitions of PA in the current research community, the primary contribution of this survey is to establish the ***first structured and unified overview of PA in large vision models***. We introduce a comprehensive taxonomy that first conquers numerous prompt-related research in large vision models into a unified scope, and then, in detail, divides them based on their distinct algorithmic designs and usages.

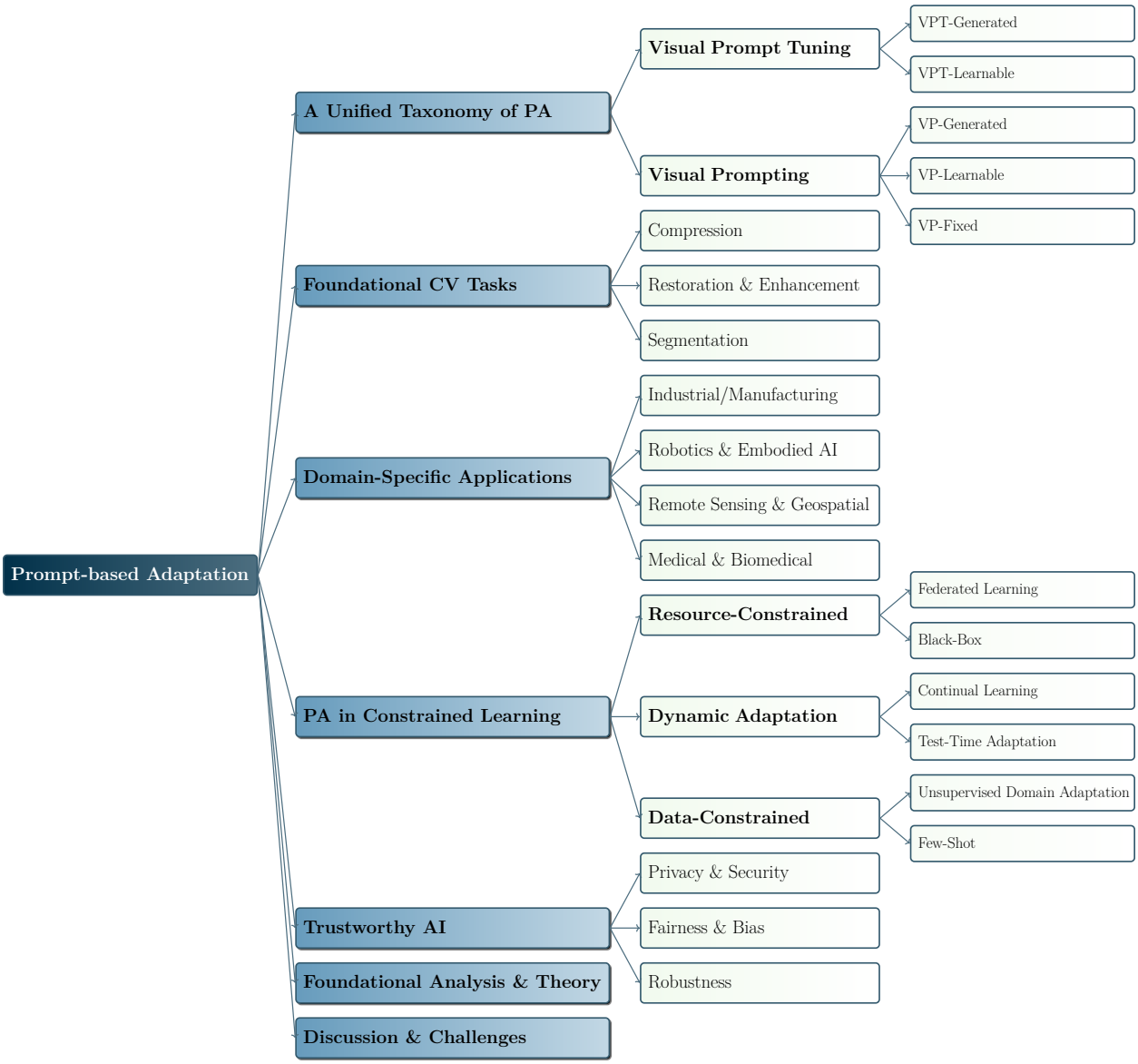

Figure 1: **A taxonomy of Prompt-based Adaptation (PA) in Large Vision Models.**

Our work is structured as follows. In §2, we begin by defining the overarching discipline of PA as the process of designing inputs at different locations to finetune a model's behavior. Within this field, we distinguish between two core paradigms in the visual domain: ❶ **Visual Prompting (VP)** and ❷ **Visual Prompt Tuning (VPT)**. In §2.2–2.3, we present the algorithmic foundations of VP and VPT, respectively, highlighting their related yet distinct perspectives toward achieving parameter efficiency. This categorization is determined by the geometric placement of prompts, distinguishing between those that modify the model's input and those that are integrated internally prior to the layer(s). In §2.4, we discuss the scopes of efficiency that PT and VPT focus on. In §3, we include PA's applications on foundational computer visions tasks, such as segmentation, restoration and enhancement, and compression. In §4, we explore the expanding applications of PA across advanced machine learning problems and in various domain-specific contexts, such as medical imaging and robotics. In §5, our survey indicates that PA successfully demonstrates effectiveness in various scenarios with optional constraints. In §6, we discuss PA *w.r.t.* trustworthy, specifically categorizes into robustness, fairness and bias mitigation, and privacy and security. In §7, we delve into the foundational analysis and theoretical underpinnings of PA. Last but not least, in §8, we discuss key challenges and identify

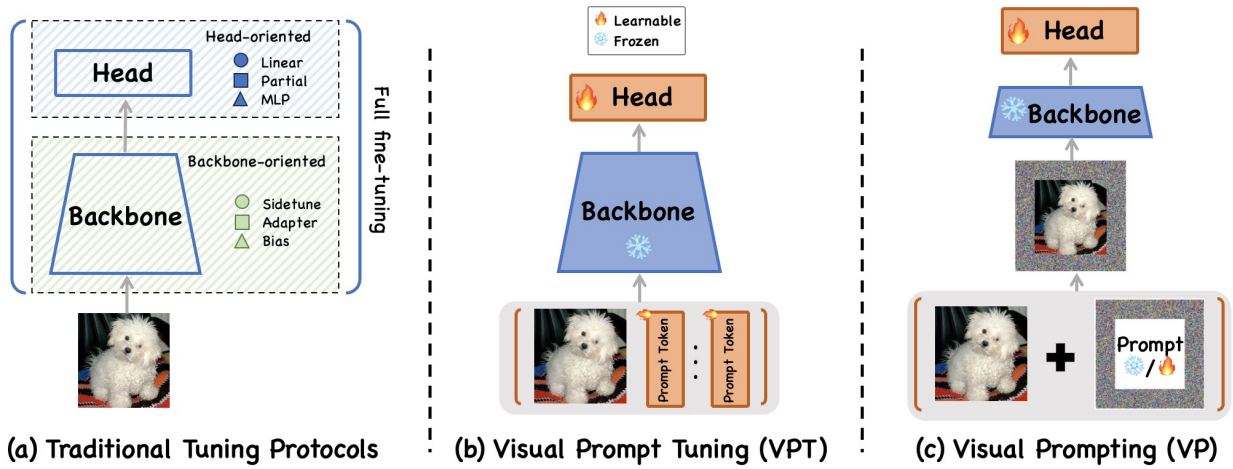

Figure 2: **Comparison of transfer learning and prompt-based adaptation methods.** (a) Current transfer learning protocols are grouped by tuning scope: Full fine-tuning, head-oriented, and backbone-oriented approaches. (b) Visual Prompt Tuning (VPT) freezes the backbone and optimizes additional prompt tokens together with the head. (c) Visual Prompting (VP) instead modifies the input space by adding prompts (which can be fixed or learnable), while keeping the backbone frozen and training a lightweight head.

PA's promising future directions. The discussion encompasses pressing issues that remain to be addressed in the PA community, including safety considerations, training and inference latency, stability, and obstacles to real-world deployment. Acknowledging that PA has already been utilized in real-world scenes, the discussions are particularly valuable for guiding future research.

Related Works. Existing surveys related to PA focus on limited scopes, as they mainly focus on multimodal or vision–language settings. For example, Wu et al. (2024d) focuses on visual prompting in MLLMs, organizing techniques around visual instructions, prompt generation, and compositional reasoning, without covering internal pixel/token injections or parameter-efficient tuning in vision encoders. (Gu et al., 2023) provides a systematic review of prompt engineering on vision–language foundation models (*e.g.*, CLIP/Flamingo/Stable Diffusion), emphasizing text-side prompts and VL pipelines rather than PA mechanisms inside vision backbones. (Lei et al., 2024) surveys prompt learning in computer vision from an AIGC-centric perspective grounded in VLMs and generative models, but do not unify methods by injection granularity or constrained-paradigm deployment. *Model Reprogramming: Resource-Efficient Cross-Domain Machine Learning* (Chen, 2022) provides an early theoretical foundation for VP by framing pixel-space transformation as a learnable input reprogramming layer for cross-domain transfer, inspiring later parameter-efficient prompting methods. More recently, (Ye et al., 2025) extends the discussion to large vision and multimodal models, tracing the evolution of VP techniques from pixel-level manipulation to foundation-level adaptation. In contrast, noticing the distinct differences on text-side and vision-side prompt-related attempts, we center on PA specifically on vision models, and propose a unified taxonomy that defines and disentangles previously ambiguous PA definitions. We further classify methods by generation mechanism (*i.e.*, learnable/generative/non-learnable) and injection granularity (*i.e.*, pixel- *vs.* token-level). Beyond methodology, we systematize constrained learning paradigms (*i.e.*, few/zero-shot, TTA, continual, black-box, forward-only, federated), consolidate domain applications (*i.e.*, medical, remote sensing, robotics, industrial), and add foundational analyses (*i.e.*, behavioral evidence and efficiency/theory), offering a deployment-oriented guidance not covered by prior surveys.

2 A Unified Taxonomy of Prompt-based Adaptation in Large-scale Vision Models

This section presents a taxonomy (see Figure 1) for prompt-based adaptation in large vision models. To avoid confusion, we separate where a prompt acts from how it is obtained. To provide a clear overview of the methods discussed, we summarize representative works for both VP and VPT in Table 1. Specifically,

a method is considered representative if it is: (1) a pioneering work that established a key paradigm; (2) a canonical example of an algorithmic sub-type; or (3) a notable variant that demonstrates the field's diversity.

2.1 Preliminary on Large-scale Vision Models

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ be an image. A frozen pre-trained vision encoder f_ϕ maps $\mathbf{x}$ to a representation that we view as a sequence (or grid) of features $Z^{(0)} \in \mathbb{R}^{T \times d}$. Here T denotes the number of "sites": patch tokens for Transformers (Dosovitskiy et al., 2021; Liu et al., 2021), spatial cells for ConvNets (He et al., 2016; Liu et al., 2022), or state steps for state-space encoders (Gu & Dao, 2023; Zhu et al., 2024b). The encoder consists of L stacked blocks and produces $Z^{(L)}$; a task head h_ω (*e.g.*, a linear classifier, detector, segmentation head) outputs the prediction $\hat{\mathbf{y}} = h_\omega(Z^{(L)})$. In prompt-based adaptation we freeze ϕ and train only prompt parameters and, optionally, ω.

2.2 Visual Prompting (VP): Input-space Prompting

VP modifies the input before tokenization/feature extraction via a prompt function $u(\,\cdot\,;\theta)$ (*e.g.*, (Bahng et al., 2022)):

$$\tilde{\mathbf{x}} \;=\; u(\mathbf{x};\theta), \qquad \hat{\mathbf{y}} \;=\; h_\omega\big(f_\phi(\tilde{\mathbf{x}})\big). \tag{1}$$

In general, VP can be further categorized into three different approaches: VP-Fixed, VP-Learned, and VP-Generated, based on how these prompts are generated:

- ***VP-Fixed*** introduces no learnable θ (*e.g.*, points/boxes/masks in interactive segmentation), thus its formulation remains identical to Eq. 1. These prompts are provided by rules or users without training. Typical forms are points, boxes, or masks for interactive segmentation (*e.g.*, SAM) (Kirillov et al., 2023) and simple visual/text hints for VLMs. These prompts are intuitive and zero-shot friendly, but their capacity is bounded by the prompt design space and the interface of the underlying model.
- ***VP-Learnable*** optimizes θ in pixel space (*e.g.*, overlays, masks, residuals, frequency cues) while keeping ϕ frozen (Bahng et al., 2022):

$$\min_{\theta,\,\omega}\; \mathbb{E}_{(\mathbf{x},\mathbf{y})}\Big[\,\mathcal{L}\big(h_\omega(f_\phi(u(\mathbf{x};\theta))),\mathbf{y}\big)\,\Big] \;+\; \lambda\,\mathcal{R}(\theta). \tag{2}$$

  The learning process can be gradient-based (white-box), query-based (zeroth-order), or driven by small auxiliary modules. Fourier- or style-based prompts improve robustness and transfer under distribution shift. For medical segmentation, Fourier Visual Prompting (FVP) and Data-Dependent Frequency Prompt (DDFP) learn frequency-domain cues that regularize features across unseen domains (Wang et al., 2023c; Yin et al., 2025). OT-VP learns a universal visual prompt for target domains by aligning distributions with Optimal Transport (OT) (Zhang et al., 2025d), showing strong source-free or test-time adaptation performance (Zhang et al., 2025d). Local-Prompt introduces spatially local input prompts to reduce false positives in few-shot OoD detection (Zeng et al., 2024a). These methods keep adaptation external to the backbone yet deliver sizable gains under domain shift. Early "model reprogramming" shows that input-space patterns can repurpose black-box models with scarce data (Tsai et al., 2020). More recent works adopt zeroth-order or gradient-free updates to learn input prompts when gradients are unavailable. This line keeps the backbone untouched and fits API-only access patterns.
  For instance-level perception in satellite imagery, RSPrompter learns input prompts that guide instance segmentation with visual foundation models (Chen et al., 2024b). Promptable instance segmentation in remote sensing (*e.g.*, Insight Any Instance; ZoRI) leverages input cues to improve generalization and zero-shot recognition across scenes and sensors (Wang et al., 2024a; Huang et al., 2025a). In hyperspectral tracking, PHTrack and SPTrack inject spectral- or similarity-based input prompts that better exploit spectral structure for target localization (Chen et al., 2024c; Guo et al., 2024). In medical image segmentation, PASS performs test-time prompting to adapt styles and shapes without retraining the backbone (Zhang et al., 2024a).
  VP-Learned offers a simple, portable handle for robustness, OoD, and cross-domain transfer. It is effective when internal states are inaccessible or when we target low-cost deployment.

- ***VP-Generated*** utilizes a small generator g_ψ to produce instance-adaptive prompts in input space (*cf.* black-box/instance-adaptive VP (Oh et al., 2023)):

$$\tilde{\mathbf{x}} \;=\; u\big(\mathbf{x},\, g_\psi(\mathbf{x})\big) \;\;=\;\; (1-\mathbf{m}) \odot \mathbf{x} \;+\; \mathbf{m} \odot \mathbf{r}_\psi(\mathbf{x}), \tag{3}$$

  where $\mathbf{m}$ is a spatial mask and $\mathbf{r}_\psi$ a synthesized residual/overlay. A canonical VP formulation casts prompting as inpainting: given masked input $(\mathbf{x}, \mathbf{m})$, predict discrete visual tokens $\hat{z}_i$ and decode to pixels (Bar et al., 2022),

$$\hat{z}_i \;=\; \arg\max_{z_i}\, p_\theta\big(z_i \mid \mathbf{x}, \mathbf{m}\big), \quad \hat{\mathbf{y}} \;=\; \mathrm{Dec}(\hat{\mathbf{z}}), \tag{4}$$

  which operationalizes the prompt as a grid of input–output examples plus the query (all in pixel space). In black-box settings, θ (or ψ) can be updated with zeroth-order, as used in black-box visual prompting (Oh et al., 2023):

$$\widehat{\nabla}_\theta \mathcal{L} \;=\; \frac{\mathcal{L}(\theta + \alpha\Delta) - \mathcal{L}(\theta - \alpha\Delta)}{2\alpha}\,\Delta, \quad \Delta \sim \{\pm 1\}^{\dim(\theta)}. \tag{5}$$

  Here, VP-Generated uses an auxiliary module to synthesize the input prompt per image (or per task), making the prompt instance-adaptive while remaining in pixel space. Typical generators include lightweight multi-layer perceptrons (MLPs) or small-scale CNNs.
  For instance, BlackVIP builds a small-scale "coordinator" network that produces input-conditioned prompts; the coordinator is optimized with zeroth-order queries to a black-box model, enabling robust transfer without accessing internals (Oh et al., 2023). This design inherits the portability of VP and the flexibility of per-instance prompting, and it matches real API constraints. Other works synthesize per-image residuals or inpainting-style regions that steer the backbone at inference. They remain external to the model and are useful when gradients are not available or when we desire a unified input-facing adaptation across many backbones. Generated VP improves flexibility over static or global patterns while preserving the advantages of input-space control. It is a practical choice for black-box or multi-backbone ecosystems where a single input adaptor must generalize widely.

Note that for VP, all operations and categorizations focus on the pixel-level prompts. For internal injection prompts (see §2.3), even if learned or generated, they are defined as VPT since they act as learnable prompts into the token/feature sequence inside the network.

2.3 Visual Prompt Tuning (VPT): Internal Prompting

VPT injects learnable prompts into the token/feature sequence while freezing the backbone (Jia et al., 2022). While the majority of research explores VPT within the Transformer-based architectures, it is actually a general solution for various deep learning backbones (*e.g.*, ResNet (He et al., 2016), ConvNeXt-B (Liu et al., 2022)). In our survey, we adopt the attention layers in ViT as the primary example to illustrate VPT.

For a ViT with L layers, patch embedding yields $X^{(0)} = [\mathbf{x}_{\mathrm{cls}}^{(0)}; \mathbf{x}_1^{(0)}; \ldots; \mathbf{x}_N^{(0)}] \in \mathbb{R}^{(1+N)\times d}$. Following (Jia et al., 2022), two variants of VPTs can be distinguished based on the number of layers into which learnable prompts are incorporated. Specifically, VPT-Shallow prepends p prompt tokens $P^{(0)} \in \mathbb{R}^{p\times d}$ only at the first layer (Jia et al., 2022):

$$Z^{(0)} \;=\; [\,\mathbf{x}_{\mathrm{cls}}^{(0)};\, P^{(0)};\, \mathbf{x}_1^{(0)}; \ldots; \mathbf{x}_N^{(0)}\,], \quad Z^{(\ell+1)} \;=\; \mathrm{Block}_\ell\big(Z^{(\ell)}\big),\; \ell = 0, \ldots, L-1. \tag{6}$$

VPT-Deep, on the other hand, uses layer-wise prompts $P^{(\ell)}$ by concatenating them at each layer's input (Jia et al., 2022):

$$Z_{\mathrm{in}}^{(\ell)} \;=\; [\,\mathbf{x}_{\mathrm{cls}}^{(\ell)};\, P^{(\ell)};\, \mathbf{x}_1^{(\ell)}; \ldots; \mathbf{x}_N^{(\ell)}\,], \quad Z^{(\ell+1)} \;=\; \mathrm{Block}_\ell\big(Z_{\mathrm{in}}^{(\ell)}\big). \tag{7}$$

Each Transformer block applies layer normalization multi-head self-attention (LN–MSA) and layer normalization multi-layer perceptron (LN–MLP) with residual connections:

$$\tilde{Z}^{(\ell)} \;=\; Z^{(\ell)} \;+\; \mathrm{MSA}\big(\mathrm{LN}(Z^{(\ell)})\big), \quad Z^{(\ell+1)} \;=\; \tilde{Z}^{(\ell)} \;+\; \mathrm{MLP}\big(\mathrm{LN}(\tilde{Z}^{(\ell)})\big). \tag{8}$$

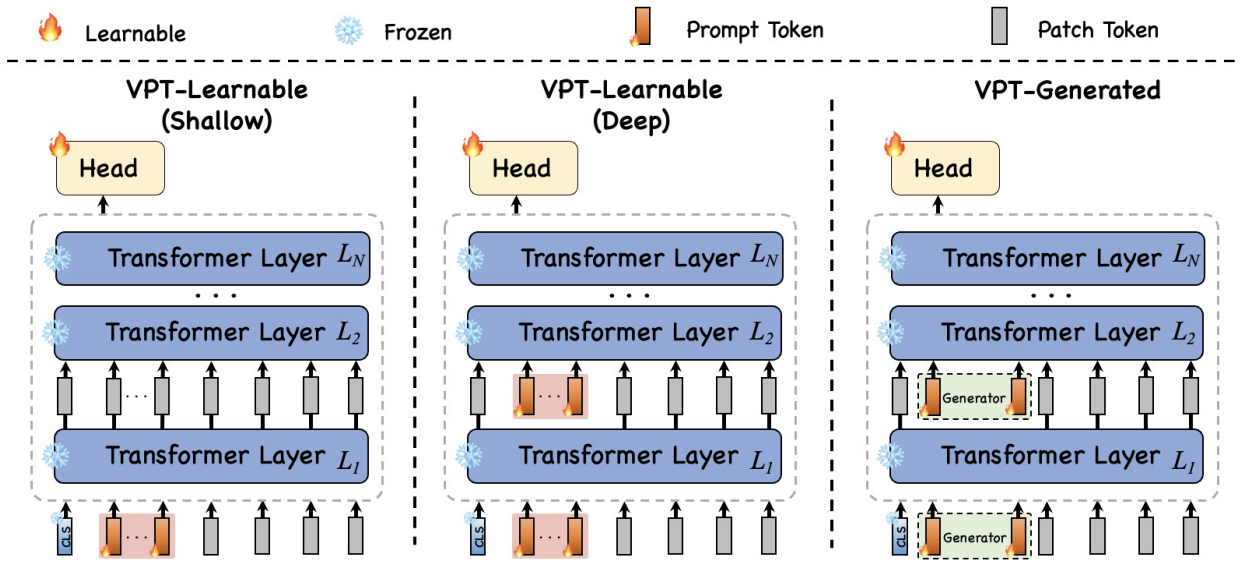

Figure 3: **Illustration of different variants of Visual Prompt Tuning (VPT, see §2.3)**. Left: **VPT-Learnable** (Shallow), where prompt tokens are only added at the first layer. Middle: **VPT-Learnable** (Deep), where prompt tokens are injected at every transformer layer. Right: **VPT-Generated**, where a generator produces instance-adaptive prompt tokens that are then inserted.

Only the prompts and head are trained (Jia et al., 2022) during finetuning:

$$\min_{\{P^{(\ell)}\},\,\omega} \mathbb{E}_{(\mathbf{x},\mathbf{y})}\Big[\mathcal{L}\big(h_\omega(\mathrm{VPT}_{\{P^{(\ell)}\}}(f_\phi,\mathbf{x})),\mathbf{y}\big)\Big], \quad \phi \text{ frozen.} \tag{9}$$

From a methodology perspective, VPT can be categorized into VPT-Learnable and VPT-Generated:

- ***VPT-Learnable*** utilizes a small number of learnable prompt tokens added to the token sequence while keeping ϕ frozen (*cf.* Eqs. 6–9). These learnable prompts are optimized by gradient descent and can be inserted only at the first layer (*i.e.*, shallow) or at every layer (*i.e.*, deep) (Jia et al., 2022). Beyond the baseline VPT (*i.e.*, here stands for Jia et al. (2022)), many variants refine what tokens encode and how they are scheduled: For long-tailed classification, LPT adds class-aware prompt tokens and a re-weighted training schedule to balance head *vs.* tail classes (Dong et al., 2023). For self-supervised ViTs, improved token initialization/regularization stabilizes adaptation and reduces the gap to FT (Yoo et al., 2023). EXPRES builds learnable "output" tokens and residual prompt tokens to better steer frozen transformers, improving VTAB/FGVC benchmark performance with modest token counts (Das et al., 2023). SA^{2}VP learns a spatially aligned 2D map of prompt tokens and adapts them across depths via cross-attention, yielding stronger transfer on dense/classification tasks (Pei et al., 2024). VFPT augments prompt tokens with Fourier components to capture frequency cues, improving robustness under distribution shift with low parameter (Zeng et al., 2024b). SPT provides design heuristics on token length, placement, and initialization that consistently lift standard VPT (Wang et al., 2024b). E^{2}VPT introduces key–value prompt tokens and pruning to cut parameters/FLOPs while retaining accuracy, scaling from shallow to deep injection (Han et al., 2023). Adaptive Prompt tunes the prompt schedule (length/placement) with simple rules or meta-updates to reduce manual search (Le et al., 2025). DA-VPT (semantic-guided) aligns the distribution of prompt tokens with class semantics via metric learning, improving generalization across tasks (Ren et al., 2025).
- ***VPT-Generated*** uses a lightweight generator (*e.g.*, an MLP or a hypernetwork) to produce **prompt tokens** conditioned on the input or task, which are then inserted shallowly or layer-wise (*cf.* Eqs. 6–7). The motivation of this design is to improve prompts' instance adaptivity and reduce the manual and rigid design of their layouts. Instance-adaptive designs generate a token set per image: DVPT for

medical analysis uses a bottleneck and cross-attention to derive sample-specific queries before emitting prompt tokens (He et al., 2025a); another DVPT variant employs a Meta-Net to produce a unique token for each image across recognition tasks (Ruan & Wang, 2023). ViaPT generates instance-aware prompt tokens for each image, enabling the model to better capture intra-class diversity while keeping the backbone frozen (Xiao et al., 2025c). Long-horizon conditioning can also drive token generation: LSPT gates information from earlier blocks to synthesize long-term spatial prompt tokens for self-supervised ViTs (Mo et al., 2024). Beyond recognition, a prompt token generator improves generative transfer in ViT-based synthesis (Sohn et al., 2023). Prompt Generation Networks (PGN) learn per-sample prompts with a tiny network; the generator itself operates in the latent/token space and integrates with frozen ViTs (Loedeman et al., 2024). Overall, the extra computational cost is the parameters and FLOPs of g_ψ; accuracy–efficiency is governed by the depth of injection (*i.e.*, shallow *vs.* deep) and the number of generated tokens.

2.4 Efficiency in Practice

The overall memory usage of fine-tuning a model can be divided into four parts: (1) *Model memory*, the storage of parameters; (2) *Activation memory*, the cache of intermediate features during the forward pass; (3) *Gradient memory*, the storage of gradients during backpropagation; and (4) *Optimizer memory*, the additional states maintained by optimizers (*e.g.*, momentum and variance in Adam). In full fine-tuning, all four components scale with the backbone size.

Efficiency of VPT. VPT freezes the backbone and updates only a small set of prompt tokens and the task head, so parameter gradients and optimizer states are allocated only for these lightweight modules. However, backpropagation must still traverse the entire backbone to compute token gradients, meaning that activation memory, which typically dominates the total GPU usage, remains largely unchanged. Thus, VPT effectively reduces the parameter and optimizer footprint but marginally alleviates activation-related memory cost (*e.g.*, full fine-tuning a vision backbone with hundreds of millions of parameters requires updating all weights, while VPT only introduces a small set of prompt tokens (*i.e.*, typically $< 0.5\%$ of parameters) that are optimized while the backbone remains frozen (Jia et al., 2022)).

While parameter-efficient approaches focus on reducing the number of trainable weights, a complementary line of work aims at minimizing activation memory during training. Recent memory-efficient fine-tuning (MEFT) (Kim et al., 2023; Simoulin et al., 2024) explores adaptive token or feature selection to avoid storing gradients for redundant activations, achieving substantial reductions in peak GPU memory with negligible performance degradation. These efforts suggest that parameter sparsity alone is probably insufficient for large-scale efficiency: activation optimization is becoming one of the key factors for scaling fine-tuning on commodity hardware.

To sum up, VPT still offers a highly efficient trade-off between accuracy and resource usage. By updating only a compact set of prompt tokens, it enables fine-tuning large-scale vision models on commodity GPUs and accelerates deployment in memory-constrained or latency-sensitive settings.

Efficiency of VP. VP operates at the input space and freezes the backbone by design. For *VP-Fixed* (*e.g.*, points/boxes/masks in SAM (Kirillov et al., 2023) or inpainting-style prompting (Bar et al., 2022)), there are *no trainable prompt parameters*, hence neither parameter gradients nor optimizer states are maintained for prompting itself; training-time memory reduces to the head (if any) and the activations required to backpropagate through the frozen backbone. For *VP-Learned* (Bahng et al., 2022) and *VP-Generated* (*e.g.*, BlackVIP (Oh et al., 2023), frequency- or distribution-aware variants such as FVP/DDFP (Wang et al., 2023c; Yin et al., 2025)), the trainable footprint remains lightweight (θ or a small generator g_ψ), so *parameter gradients and optimizer states are allocated only for these modules and the task head*, not for the backbone. However, as with VPT, backpropagation must still traverse the entire backbone to compute gradients *w.r.t.* θ or g_ψ, which implies that *activation memory largely remains*, typically the dominant component of peak GPU usage during training. Thus, VP effectively minimizes parameter/optimizer overhead (and eliminates it entirely in VP-Fixed), but only marginally alleviates activation-related cost.

Table 1: **Representative Visual Prompting (VP) and Visual Prompt Tuning (VPT) methods.**

	Method	Venue	Year	Sub-type	Prompt Space
VP	VPI (Bar et al., 2022)	NeurIPS	2022	Generated	Pixel
	LabelMap (Chen et al., 2023)	CVPR	2023	Learnable	Pixel
	SAM (Kirillov et al., 2023)	CVPR	2023	Fixed	Pixel
	DAM-VP (Huang et al., 2023b)	CVPR	2023	Generated	Pixel
	Fg-VP (Yang et al., 2023b)	NeurIPS	2023	Fixed	Pixel
	SSMask (Liu et al., 2024)	ICML	2024	Learnable	Pixel
	InsVP (Zheng et al., 2024)	ACM MM	2024	Generated	Pixel
	PixelVP (Sun et al., 2024b)	TMLR	2024	Learnable	Pixel
	BayesVRP (Zhao et al., 2024b)	NeurIPS	2024	Learnable	Pixel
	AutoPrompt (Ravi et al., 2024)	arXiv	2024	Generated	Pixel
	AttrVP (Chen & Wang, 2025)	ICLR	2025	Learnable	Pixel
	LoR-VP (Jin et al., 2025)	ICLR	2025	Learnable	Pixel
VPT	VPT (Jia et al., 2022)	ECCV	2022	Learnable	Token
	LPT (Dong et al., 2023)	ICLR	2023	Learnable	Token
	EXPRES (Das et al., 2023)	CVPR	2023	Learnable	Token
	SSL-VPT (Yoo et al., 2023)	ICML	2023	Learnable	Token
	E^2VPT (Han et al., 2023)	ICCV	2023	Learnable	Token
	VPT-Gen (Sohn et al., 2023)	CVPR	2023	Generated	Token
	SA^2VP (Pei et al., 2024)	AAAI	2024	Learnable	Token
	RePrompt (Wang et al., 2024b)	arXiv	2024	Learnable	Token
	LSPT (Mo et al., 2024)	AAAI	2024	Generated	Token
	VFPT (Zeng et al., 2024b)	NeurIPS	2024	Generated	Token
	AdaPrompt (Le et al., 2025)	arXiv	2025	Learnable	Token
	SG-VPT (Ren et al., 2025)	CVPR	2025	Learnable	Token
	DVPT (He et al., 2025a)	NN	2025	Generated	Token
	ViaPT (Xiao et al., 2025c)	ACM MM	2024	Generated	Token
	SPT (Wang et al., 2024b)	ICML	2024	Learnable	Token

In terms of practicality, VP offers two further efficiency advantages. First, VP-Fixed supports training-free or head-only adaptation, which *removes* prompt-side gradients and optimizer states by construction; the remaining memory is mostly due to activations through the frozen backbone and the small head, enabling extremely lightweight deployment settings. Second, VP is *black-box friendly*: gradient-free/zeroth-order optimization (as in BlackVIP (Oh et al., 2023)) avoids storing parameter gradients altogether, trading queries for memory and widening the feasibility on API-only backbones. During inference, composing prompts in the pixel space introduces negligible computational overhead (*e.g.*, border/overlay composition) relative to the main forward through the backbone; the runtime and memory are therefore dominated by the frozen encoder pass.

To sum up, VP provides a complementary efficiency profile to VPT: it *minimizes* parameter and optimizer states at the prompt side (to zero in VP-Fixed), preserves a frozen backbone, and enables black-box or training-free use cases, while leaving activation memory largely unchanged during training. This makes VP a practical choice for commodity hardware and latency-sensitive deployment, especially when API-only access or user-interactive prompting is required.

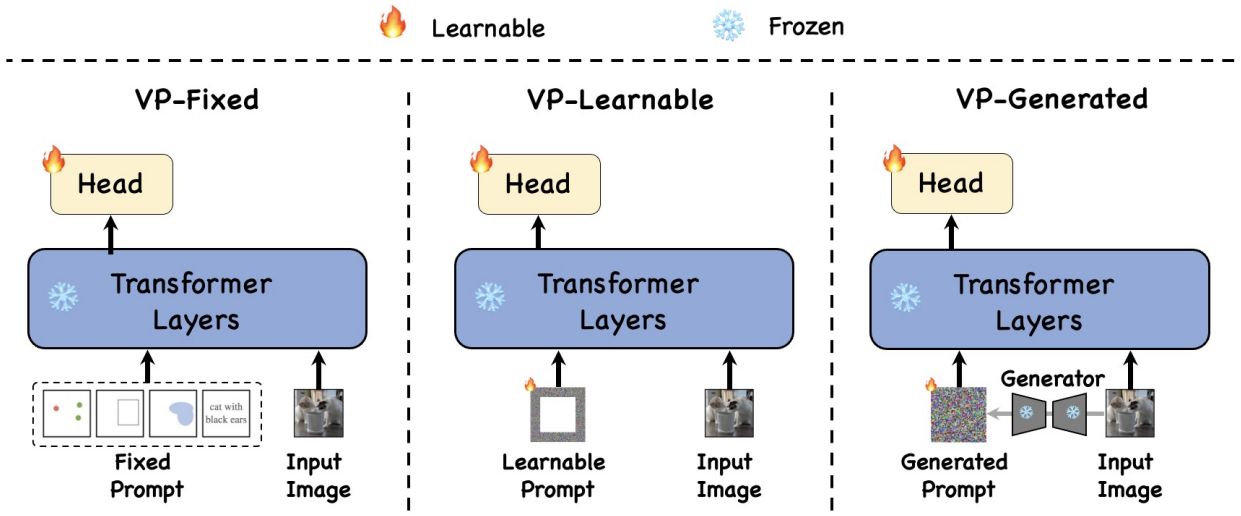

Figure 4: **Illustration of Visual Prompting (VP) variants (see §2.2).** (Left) **VP-Fixed**: prompts are predefined (e.g., boxes, points, text hints) and directly attached to the input without training. (Middle) **VP-Learned**: learnable prompts in the pixel space are optimized jointly with the task head while the backbone remains frozen. (Right) **VP-Generated**: a generator network produces instance-adaptive prompts for each input image, offering higher flexibility.

2.5 Comparative Analysis of Generation Mechanisms

Beyond efficiency trade-off, the selection among fixed, learnable, and generated prompting mechanisms is governed by the specific requirements of adaptation depth, data availability, and deployment constraints.

Fixed Prompts. This category, exemplified by the explicit points or boxes in SAM (Kirillov et al., 2023) and predefined templates in visual-language models, excels in *interactivity* and *interpretability*. Since these prompts are immutable during inference, they require no gradient updates, making them the only viable option for strictly black-box adaptation scenarios where model weights are inaccessible. However, their performance is strictly upper-bounded by the inherent zero-shot capability of the frozen backbone. They struggle to adapt to systematic domain shifts (*e.g.*, medical spectral bands) where the pre-trained feature space is misaligned with the target distribution, as they lack the capacity to recalibrate internal representations.

Learnable Prompts. Optimizing continuous vectors, as seen in VPT (Jia et al., 2022) and VP (Bahng et al., 2022), represents the standard paradigm for domain adaptation. By leveraging white-box gradient access, learnable prompts can bridge significant domain gaps (*e.g.*, ImageNet to Remote Sensing) by overfitting a lightweight set of parameters to the target distribution. Nevertheless, a critical limitation is their *static nature*: once trained, the prompt remains identical for all input instances. This "dataset-level" optimization overlooks intra-class variations and instance-specific nuances, potentially limiting performance in highly diverse or open-world scenarios where dynamic adjustment is required.

Generated Prompts. To address the static limitation of learnable prompts, generated mechanisms (*e.g.*, IDPT (Zha et al., 2023), DAM-VP (Huang et al., 2023b)) introduce *instance-awareness*. By conditioning the prompt generation on individual input images via a lightweight network, these methods allow the model to dynamically adapt to varying lighting, scale, or occlusion conditions on a per-sample basis. This dynamic capability typically yields superior generalization and robustness compared to static learnable prompts. The trade-off, however, lies in the increased inference latency and optimization complexity, as the generator introduces an additional computational overhead and must be carefully regularized to prevent overfitting to the training distribution.

To sum up, fixed prompts offer the highest deployment flexibility; learnable prompts provide a strong baseline for stable domain transfer; and generated prompts offer the highest theoretical ceiling for complex, instance-varying tasks at the cost of computational complexity.

3 Foundational Computer Vision Tasks

We first introduce PA in foundational computer vision tasks (*i.e.*, segmentation (see §3.1), restoration and enhancement (see §3.2), compression (see §3.3), multi-modality (see §3.4)), aiming to provide a comprehensive guideline on how prompt mechanisms can bridge pre-trained large-scale vision models with downstream visual recognition and understanding.

3.1 Segmentation

PA methods are now being applied across a wide range of segmentation paradigms. In continual segmentation, *ECLIPSE* maintains a frozen backbone and adds visual prompts to update panoptic segmentation models, thereby mitigating catastrophic forgetting and ensuring stable plasticity over long task sequences (Kim et al., 2024a). For multimodal scenarios, *DPLNet* designs dual prompts for RGB and auxiliary modalities, fusing them with a lightweight module to minimize the number of trainable parameters (Dong et al., 2024a). *GoPT* further group features into semantic clusters and inject prompts on a per-group basis, which improves cross-modal alignment using less than 1% of the total model weights for training (He, 2024).

Few-shot segmentation has benefited from class- and instance-aware prompting. *PAT* constructs prompts by transferring semantic cues from base to novel classes and by generating part-aware masks, an approach that enhances adaptation to new categories (Bi et al., 2024). *Hossain et al.* employ multi-scale prompting and causal attention between base and novel classes to strike a balance between generalization to new data and retention of prior knowledge (Hossain et al., 2024). Beyond the RGB domain, *Xu et al.* augment CLIP encoders with spectral prompts and utilize a spectral-guided decoder to improve pixel-level adaptation for unseen classes (Xu et al., 2024a). Prompting has also proven useful for low-level structural segmentation. *EVP* formulates explicit prompts by combining features from frozen patch embeddings with high-frequency components; this method unifies camouflaged object, forgery, shadow, and defocus-blur segmentation within a single framework (Liu et al., 2023c). To address domain shifts in annotation style or image statistics, domain-adaptive prompting for SAM (*DAPSAM*) demonstrates that even interactive models can be steered by prompts learned in a new domain without altering the main model weights (Zhang et al., 2024d). Finally, VPT-style token prompts can be spatially aligned or implemented in a layer-wise fashion to better suit dense prediction tasks; SA^2VP, for example, aligns a 2D map of prompt tokens across network depths, improving both semantic and panoptic segmentation with few additional parameters (Pei et al., 2024).

3.2 Restoration and Enhancement

Image restoration must contend with a wide variety of degradations. PA offers a lightweight and flexible mechanism for encoding degradation-specific cues without re-engineering the entire network architecture. For all-in-one restoration, *PromptIR* injects degradation-aware prompts that guide a single backbone model across denoising, dehazing, and deraining tasks (Potlapalli et al., 2023). *PromptRestorer* extends this concept by extracting raw degradation features and processing them through a dedicated prompting branch with global–local modulation and gated propagation. This approach stabilizes training and elevates restoration quality across four distinct tasks (Wang et al., 2023a).

Frequency-aware prompts are effective for recovering fine details that spatial-only cues might miss. *FPro* decomposes features into frequency bands and utilizes dual prompt blocks to modulate low and high frequencies independently, demonstrating improved performance on five restoration tasks (Zhou et al., 2024b). Diffusion-based approaches can also be guided by prompts: *Diff-Restorer* leverages priors from Stable Diffusion and employs visual prompts to tackle a range of degradations within a unified framework (Zhang et al., 2024e). For compressed images, *PromptCIR* learns distortion-aware prompts that adapt to varying compression levels and artifacts, enhancing both fidelity and perceptual metrics (Li et al., 2024a). Beyond these methods, "ingredient-oriented" prompting encodes specific task factors as "ingredients," allowing a single model to dynamically switch behaviors as needed (Gao et al., 2023). In summary, these results point to a

consistent pattern: by learning a small set of prompts that encapsulate what varies across conditions—be it the degradation type, frequency band, or compression level—it is possible to leave the core backbone frozen. This strategy keeps training computationally efficient while delivering substantial gains in restoration quality in practice.

3.3 Compression

PA is increasingly being applied to image and video compression, primarily to enhance adaptability (*e.g.*, enabling variable bit rates, region-of-interest (ROI) control, or user-defined quality preferences), without retraining large codec models end-to-end. One line of research conditions Transformer-based learned image compression with prompts, allowing a single model to cover multiple operating points. For variable-rate coding with ROI support, Kao *et al.* generate content-adaptive *prompt tokens* derived from the input image, an ROI mask, and a target rate. These tokens are then fed into the codec's Transformer blocks, effectively decoupling the decision of *what to preserve* (the ROI) from *how much to spend* (the bit rate) within a unified model (Kao et al., 2023). The same group further conditions the codec on user-selected *quality objectives* (*e.g.*, different perceptual metrics) via prompt tokens produced by a lightweight prompt generator, again avoiding the need for multiple specialized checkpoints (Kao et al., 2023).

Another approach adapts visual prompt tuning to modulate pretrained Transformer codecs. Qin *et al.* propose a layer-adaptive prompt module that injects prompts into both the encoder and decoder to steer attention and bit allocation across different rates. This method reduces parameter storage and data requirements while matching the rate–distortion performance of separately trained models (Qin et al., 2024). At a conceptual level, recent work on "rate–distortion–cognition (RDC)" argues that learned codecs should expose controllable parameters that balance human-perceived quality, machine utility, and bitrate; prompts serve as a natural mechanism for implementing such controls and conveying side information (Jia et al., 2024).

In the video domain, two complementary applications have emerged. First, prompts can drive *semantic* compression by aligning a compact representation with high-level tasks. Free-VSC learns prompts specific to vision foundation models to guide unsupervised video semantic compression without labels, optimizing the compressed stream for downstream understanding by vision–language models (Tian et al., 2024b). Similarly, SMC++ demonstrates that masked learning with prompt-like conditioning improves semantic coding pipelines and simplifies their engineering at scale (Tian et al., 2024a). Second, prompts can be applied on the *consumer* side: Compressed Video Prompt Tuning adapts models trained on raw video to compressed-domain inputs (*e.g.*, motion vectors, residuals) by re-parameterizing them into conditional prompts, which enhances recognition performance without retraining on raw pixels (Li et al., 2023a).

Prompted compression is also appearing at the interface of downstream tasks. UCIP utilizes dynamic prompts for universal compressed image super-resolution, improving restoration quality across different codecs and bitrates (Zhang et al., 2024b). For multimodal question answering, IQViC introduces a question-adaptive visual compressor that uses in-context text to determine *what content to preserve* before transmission, effectively transforming prompts into rate–content selectors (Yamao et al., 2024). Collectively, these results indicate a practical direction: prompts can function as a lightweight control plane for learned codecs (*i.e.*, dictating the rate, region, quality objective, or task utility), while the computationally intensive backbone remains frozen.

3.4 Multi-modal Tasks

Prompting is a cornerstone of interaction with multi-modal models, most notably with vision–language models (VLMs) such as CLIP (Radford et al., 2021). Early and foundational work largely explored *text prompts* (often called "prompt engineering") for recognition and retrieval. For example, *CoOp* learns continuous context vectors as class prompts on the text branch to improve few-shot classification (Zhou et al., 2022b); *CoCoOp* makes the learned text prompt conditional on the input image to generalize better across domains (Zhou et al., 2022a); and *MaPLe* jointly learns prompts on both the text and image branches for stronger cross-modal alignment (Khattak et al., 2023). Beyond prompt embeddings, other techniques have been explored: *CLIP-Adapter* adds a lightweight residual adapter to better fuse visual features with the text

classifier (Gao et al., 2024), while *Tip-Adapter* uses cached features to build a classifier from class prototypes, avoiding full fine-tuning (Zhang et al., 2022). Similarly, open-vocabulary detection and segmentation use text queries as prompts (*e.g.*, *GLIP* (Li et al., 2022b), *OWL-ViT* (Minderer et al., 2022)), and generative models are steered via textual prompts or learned text tokens (*e.g.*, textual inversion for diffusion models (Gal et al., 2023)).

More recently, the focus has shifted from text-side engineering to leveraging *visual prompts* to directly guide Multimodal Large Language Models (MLLMs). This paradigm allows for more intuitive and fine-grained control over model behavior. Users can provide explicit visual instructions through methods like free-form drawing (Lin et al., 2024b), applying a set of marks or points to ground the model's attention (Yang et al., 2023a), or using one image as an exemplar to manipulate another (Yang et al., 2024b). Research in this area also explores how to best design these prompts for specific tasks, such as fine-grained control in image segmentation (Liu et al., 2023b; Yang et al., 2023b) or enabling new paradigms for open-vocabulary detection (Zhang et al., 2023; Wu et al., 2025b). Beyond user-provided inputs, other works focus on automating the creation of visual prompts through cross-modal optimization (Ravi et al., 2024), learning them in a training-free manner (Wu et al., 2024b), or jointly optimizing them with text prompts (Jiang et al., 2024). The transferability of these visual prompts (Jeong et al., 2025) and the integration of external knowledge (Lin et al., 2024c) are also active areas of investigation, highlighting the rapid growth of vision-side prompting in the multi-modal space.

Scope of this Survey. The methods above illustrate a rich and evolving landscape of multi-modal prompting. While we acknowledge the importance of both text-side and vision-side prompting for MLLMs, our survey's primary focus remains on the *internal mechanisms* of prompt-based adaptation (PA) on vision-centric backbones. VP and VPT, as defined in this survey, represent the core algorithmic primitives for adapting the vision encoder itself. We discuss multi-modal works to provide essential context and to clarify the distinction between adapting the VL interface versus adapting the vision pathway, which is our central theme.

4 Domain-Specific Applications

While PA has demonstrated remarkable success in conventional computer vision tasks, its full impact is best revealed through practical deployments across a range of heterogeneous, real-world domains. In this section, we provide a systematic overview of how PA is being adapted to diverse applied settings. We organize existing applications based on the unique demands and intrinsic challenges of the target environment. In the following subsections, we organize our discussion by application domain rather than strictly separating methods into VP and VPT categories. Since the adoption of these paradigms is not uniformly balanced across all fields; some domains currently favor one approach over the other. Therefore, while we comprehensively discuss how PA addresses domain-specific challenges, we refer readers to Table 2 and Table 3 for comprehensive method-level classifications of VP and VPT applications, respectively.

Current domain-specific uses of PA can be categorized into four major areas: medical and biomedical imaging (see §4.1), remote sensing and geospatial analysis (see §4.2), robotics and embodied AI (see §4.3), industrial inspection and manufacturing (see §4.4), autonomous driving and advanced driver-assistance system (ADAS) (see §4.5), 3D Point Clouds and LiDAR (see §4.6), video understanding and temporal perception (see §4.7), and underwater and adverse Environments (see §4.8). These categories are defined not by PA architectural variations but rather by the distinct operational contexts and domain-specific challenges they present.

4.1 Medical and Biomedical Imaging

Medical and biomedical imaging represents a critical domain where PA has demonstrated compelling utility, driven by the fundamental need for data efficiency, cross-modality generalization, and interpretability. PA offers a modular approach for adapting large-scale vision models to segmentation, classification, and reporting tasks across various imaging modalities (*e.g.*, CT, MRI, X-ray, histopathology), which frequently operate under limited supervision (*i.e.*, data scarcity) and strict clinical constraints (Xiao et al., 2024; 2025e; Wei et al., 2025). A prominent direction is the adaptation of vision foundation models, such as the *Segment Anything Model (SAM)* (Kirillov et al., 2023), to medical segmentation tasks via learnable visual prompts.

Works like *Customized SAM* and *3D SAM-Adapter* extend SAM to radiological and volumetric datasets by injecting 2D or 3D spatial prompts that encode lesion or organ priors (Zhang & Liu, 2023; Gong et al., 2024a). *Ma-SAM* further refines this paradigm by introducing a modality-agnostic prompt encoder capable of handling multimodal volumetric data through the joint optimization of spatial and semantic cues (Chen et al., 2024a). These designs exploit the structure-preserving benefits of pixel-based prompting while leveraging the semantic scalability of foundation models like SAM.

Beyond segmentation, VPT has also been applied to clinical report generation and multimodal reasoning. *PromptMRG* leverages diagnosis-driven prompts to align imaging features with clinical report templates, thereby improving factual alignment and coherence in generation (Jin et al., 2024). Concurrently, *Biomed-DPT* employs dual-modality prompt tuning to bridge vision-language pretraining with medical text, enabling few-shot adaptation for diverse downstream tasks such as classification, grounding, and captioning (Peng et al., 2025). This line of work highlights how prompt design can function as a conduit for clinical knowledge integration. To address distribution shifts across devices, sites, or patient populations, VPT can be applied for domain generalization and source-free adaptation. For instance, *FVP* and *DDFP* utilize frequency-domain prompts to regularize representations across domains, demonstrating strong robustness under unseen test distributions (Wang et al., 2023c; Yin et al., 2025). *ProSFDA* integrates prompt learning into source-free domain adaptation pipelines, achieving improved alignment without requiring access to source data at test time (Hu et al., 2022).

Finally, real-world deployments in federated and privacy-sensitive settings have inspired task-specific prompting strategies. *FedLPPA* proposes personalized prompt learning across decentralized clients for weakly-supervised segmentation, while *PASS* applies test-time visual prompting to adapt styles and shape priors under distribution drift (Lin et al., 2024a; Zhang et al., 2024a). SafeTriage leverages prompt-based anonymization to preserve identity protection in facial video triage while retaining task relevance (Savic & Zhao, 2023). Meanwhile, recent benchmarking efforts such as *A Real-World Dataset* provide standardized datasets for evaluating foundation model adaptation across real-world hospitals and imaging systems (Wang et al., 2023b). Together, spanning learnable tokens, pixel-space injection, and cross-modal generation, these advances reveal how VPT methods are actively reshaping medical imaging workflows. By offering a parameter-efficient, interpretable, and context-aware adaptation mechanism, VPT is poised to bridge the gap between foundation models and the safety-critical demands of clinical AI systems.

In the medical domain, the choice between VP and VPT is largely dictated by the task modality and interaction requirements. VP, particularly in its fixed and generated forms, dominates segmentation tasks. This preference stems from the inherent compatibility between pixel-space prompts (*e.g.*, points, boxes) and clinical workflows, where human-in-the-loop interaction is required to localize lesions using foundation models like SAM. Conversely, VPT prevails in high-level reasoning tasks such as report generation and disease classification. By injecting learnable tokens directly into the feature space, VPT facilitates the integration of specialized clinical knowledge and multi-modal alignment (*e.g.*, image-to-text) while maintaining the structural integrity of the frozen backbone, a critical advantage for data-efficient adaptation in resource-constrained or federated medical environments.

4.2 Remote Sensing and Geospatial Analysis

Remote sensing and geospatial analysis present unique challenges, such as domain heterogeneity, extreme class imbalance, and spectral complexity (Cheng et al., 2020). Consequently, PA has been increasingly explored as a mechanism to efficiently adapt foundation models for tasks such as segmentation, retrieval, and change detection, particularly under data-scarce or multi-modal conditions.

A central theme in this area is the adaptation of foundation models for object detection and instance segmentation. RSPrompter, for instance, utilizes task-specific visual prompts to guide instance segmentation in satellite imagery, improving precision and instance separation without full model retraining (Chen et al., 2024b). Insight Any Instance and ZoRI extend this concept to generalizable and zero-shot segmentation tasks, where instance-level visual cues are injected via learnable prompts to enhance detection robustness across diverse scenes and sensors (Li et al., 2025f; Huang et al., 2025a; Xiao et al., 2025b; Wang et al., 2025b). Prompting-based approaches have also proven beneficial for temporal reasoning tasks such as change detec-

tion and captioning. *A Decoupling Paradigm* employs prompt tuning to disentangle static and dynamic scene features, enabling the accurate description of semantic changes over time (Liu et al., 2023a). Similarly, VPT-enhanced CLIP has been used to guide bi-temporal change detectors with learned priors that highlight expected spatial transformations, thereby improving generalization under noisy labels (Liu et al., 2025c; Ji et al., 2025). Cross-modal retrieval represents another important use case where prompts serve to align vision and language spaces for effective geospatial understanding. Methods like *Parameter-efficient transfer* and *RLita* explore lightweight, prompt-based fine-tuning for satellite-to-caption and caption-to-region alignment, outperforming full model tuning while maintaining modality consistency (Yuan et al., 2023; Zhang et al., 2025b). In the context of few-shot classification, works such as *MVP* and the *Few-shot Survey* demonstrate that VPT can encode spatial priors or domain-specific knowledge, thereby improving performance on unseen classes and datasets with minimal supervision (Zhu et al., 2024a; Qiu et al., 2024). For broader foundation model adaptation, methods like *UPETU* and *LayerLink* introduce parameter-efficient prompt encoders capable of tuning pretrained vision backbones for scene classification and detection without requiring access to the full training data or GPU-intensive backpropagation (Dong et al., 2024b; Zhu et al., 2025). These efforts underscore the scalability of prompt-based fine-tuning in real-world Earth observation.

Finally, hyperspectral tracking and classification have recently gained attention through spectral-aware prompting. *PHTrack* and *SPTrack* integrate domain-specific spectral similarity prompts or spatial priors to guide transformer backbones, enabling them to better utilize rich spectral information in hyperspectral video and image tracking tasks (Chen et al., 2024c; Guo et al., 2024). In summary, these diverse applications demonstrate the utility of VPT for supporting flexible and modular model adaptation in remote sensing. It bridges data modality gaps, enhances domain robustness, and reduces computational overhead, positioning it as a promising paradigm for scalable Earth intelligence systems.

Analyzing the adaptation strategies in remote sensing reveals a clear division of labor driven by the unique characteristics of geospatial data. VP has become the go-to strategy for precision-oriented tasks, such as instance segmentation and object tracking (Chen et al., 2024b; Guo et al., 2024). The rationale is straightforward: satellite imagery is notoriously dense and cluttered; explicit pixel-level prompts (whether spatial or spectral) act as strong guidance signals, helping frozen models isolate specific targets amidst complex backgrounds where generic features might fail. In contrast, VPT finds its niche in high-level semantic tasks like scene classification and cross-modal retrieval (Zhang et al., 2025b; Zhu et al., 2024a). Given the massive visual gap between standard ground-level pre-training data (*e.g.*, ImageNet) and the "bird's-eye" view of remote sensing, injecting learnable tokens deep into the network architecture proves to be a more effective way to realign the global feature space for this distinct domain, all while bypassing the prohibitive cost of full-parameter retraining.

4.3 Robotics and Embodied AI

Robotics and embodied AI represent a rapidly evolving frontier where PA plays a critical role in bridging large-scale foundation models with the complex demands of 3D perception, motion reasoning, and embodied interaction. In the context of 3D understanding, prompts serve as a lightweight mechanism for adapting 2D pretrained models to point cloud data. Methods such as PointCLIP V2 Zhu et al. (2023) and CLIP2Point Huang et al. (2023c) align CLIP embeddings with point cloud features through contrastive prompting, while P2P Wang et al. (2022d) introduces point-to-pixel prompting that enables the direct transfer of pretrained vision-language knowledge to 3D point clouds. Further research from GAPrompt Ai et al. (2025) and PointLoRA Wang et al. (2025a) demonstrates that geometry-aware or low-rank prompts can outperform full fine-tuning in open-world and few-shot tasks. Techniques such as instance-aware dynamic prompts Zha et al. (2023) and other parameter-efficient methods like Point-PEFT Tang et al. (2024a) expand the utility of prompts for task specialization across various 3D recognition benchmarks. For embodied interaction, architectures like ShapeLLM Qi et al. (2024) and Any2Point Tang et al. (2024b) extend PA frameworks to support multi-modal grounding and affordance-based reasoning for manipulation tasks. Canonical shape prompting facilitates few-shot, class-incremental 3D learning by generating view-invariant prototypes Cheraghian et al. (2024). In the unified driving perception, visual exemplar-driven task prompting enables label-efficient scene parsing across heterogeneous sensor modalities Liang et al. (2023).

Within the video domain, prompt tuning has been shown to support temporal generalization and class-incremental learning. Space-Time Prompting introduces dynamic temporal prompts for continual video recognition Pei et al. (2023), while SimDA integrates diffusion models with prompt-driven adapters for efficient video generation Xing et al. (2024). Additionally, Zeroi2V proposes a zero-cost adaptation scheme to repurpose pretrained image transformers for video understanding, highlighting the synergy between prompt-based temporal transfer and frozen backbones in embodied AI scenarios Li et al. (2024b). Collectively, these studies illustrate that VPT serves as a flexible and parameter-efficient interface for robotics and embodied AI applications. Prompts can be used to condition models on structural priors (*e.g.*, shape, motion), guide modality fusion, and support continual adaptation in dynamic, real-world environments. As robotics increasingly relies on multi-modal and 3D inputs, prompt-based learning offers a scalable pathway for bridging the gap between general-purpose foundation models and specialized embodied intelligence.

The deployment of PA in robotics reflects a strategic effort to bridge the dimensional gap between 2D pre-training and 3D physical reality. VP, particularly via input-level projection or decoration, serves as the primary bridge for adapting vision-language models to point clouds (Wang et al., 2022d; Zhu et al., 2023). By reformatting sparse 3D data into 2D-compatible views, VP allows robotic systems to exploit the vast, open-world semantic knowledge of models like CLIP without the prohibitive cost of curating massive 3D datasets. Conversely, VPT becomes indispensable when tasks demand deep physical reasoning, such as grasping geometry or tracking temporal motion (Ai et al., 2025; Pei et al., 2023). In these scenarios, learnable tokens act as carriers for structural priors, effectively injecting the missing spatio-temporal context into the backbone that static image pre-training inherently lacks.

4.4 Industrial Inspection and Manufacturing

PA is gaining traction in high-precision manufacturing for tasks such as defect detection, anomaly segmentation, and visual quality control. A common paradigm involves maintaining a frozen backbone and using prompts to steer powerful foundation models, such as SAM or CLIP, toward product-specific visual cues with minimal or no additional supervision. At the pixel level, several works adapt the Segment Anything Model (SAM) to industrial defects using promptable masks. An unsupervised pipeline for laser-based additive manufacturing generates pseudo-labels and region masks to localize porosity without manual annotation, demonstrating that promptable segmentation is viable even with noisy factory data (Era et al., 2023). SAA+ proposes *hybrid prompt regularization* to segment anomalies without any training, improving zero-shot transfer by stabilizing how prompts guide the model (Cao et al., 2023). Self-Perception Tuning (SPT) further incorporates a lightweight self-perception branch to refine SAM's masks at test time, enhancing industrial anomaly segmentation while keeping the core model frozen (Yang et al., 2025). Extending beyond per-image prompts, SAID introduces *scene prompts*, enabling a single model to segment diverse industrial defects under varying lighting and background conditions (Huang et al., 2025b). Another line of research combines CLIP and SAM, allowing semantic cues from CLIP to serve as spatial prompts for SAM. ClipSAM aligns textual and image-based cues in CLIP and uses them to constrain SAM's masks, yielding strong zero-shot anomaly segmentation performance on benchmarks like MVTec-AD and VisA (Li et al., 2025e). A two-stage CLIP–SAM framework similarly leverages CLIP for coarse anomaly localization before refining the results with SAM, thereby boosting segmentation quality on industrial datasets (Hou et al., 2024). Additionally, image-aware prompt generators can synthesize dynamic prompts on a per-sample basis to better adapt to fine-grained surfaces and textures prior to segmentation (Zhang et al., 2025c). Prompted anomaly detection with vision–language backbones is also an active area in industrial quality assurance. WinCLIP demonstrates that carefully designed prompt ensembles and windowed features can elevate zero- and few-shot anomaly classification and segmentation (Jeong et al., 2023). AnomalyCLIP learns *object-agnostic* text prompts that generalize across different product types for zero-shot detection (Zhou et al., 2024a). VCP-CLIP injects *visual context prompts* into the text encoder of CLIP, which reduces the need for product-specific prompt engineering and improves zero-shot anomaly segmentation across numerous real-world factory datasets (Qu et al., 2024). These methods are parameter-efficient and particularly practical for scenarios where only API access to the model is available.

While less common, token-level Visual Prompt Tuning (VPT) for industrial inspection is an emerging direction. For instance, masked prompt tuning atop self-supervised features has been explored to enhance surface

defect inspection with minimal labels, showing that small sets of learnable prompts can match or exceed the performance of full fine-tuning under stringent memory constraints (Wu et al., 2025a). Overall, the trajectory is clear: prompt-based adapters (*i.e.*, whether in the pixel space (VP) or via lightweight prompt modules), offer a simple and deployable pathway for integrating foundation models into production quality assurance pipelines with low annotation cost and rapid iteration.

The industrial landscape presents a unique "rare event" challenge: defective samples are scarce, making traditional supervised training difficult. Consequently, VP has emerged as the dominant paradigm, particularly for zero-shot and few-shot anomaly detection (Jeong et al., 2023; Cao et al., 2023). By leveraging the open-vocabulary capabilities of CLIP or the zero-shot segmentation power of SAM, pixel-level or text-guided prompts allow systems to define "normality" and localize outliers without requiring exhaustive defect datasets. On the other hand, VPT is carving out a niche in surface inspection tasks involving complex textures (Wu et al., 2025a). Here, injecting learnable tokens helps fine-tune the backbone's sensitivity to subtle material variations (*e.g.*, scratches on metal versus fabric) that generic pre-trained features might overlook, offering a parameter-efficient alternative to full retraining in memory-constrained factory environments.

4.5 Autonomous Driving and Advanced Driver-Assistance System (ADAS)

The domain of autonomous driving demands exceptional robustness against severe domain shifts, such as variations in lighting (night *vs.* day), adverse weather conditions (rain and fog), sensor artifacts (lens blur), and changes in geographic location or hardware configurations. In this context, PA has emerged as a practical methodology for enhancing the performance of large vision models without necessitating complete backbone retraining. At the pixel level, visual prompting has been utilized to guide segmentation models in adverse weather conditions. Differentiable implicit prompts can be optimized end-to-end to improve road parsing on Cityscapes and related benchmarks under fog, rain, and low light (Kalwar et al., 2023). For scenarios where strong segmentation priors are available, prompting foundation models (*e.g.*, SAM-style pipelines) has proven beneficial; recent work shows that semantic prompts can transfer from Cityscapes to ACDC, Dark Zurich, and other adverse-condition datasets using simple input-level designs (Wang et al., 2024d). Furthermore, learning the prompts themselves, rather than relying on fixed prompt types, serves to close the performance gap with full fine-tuning while maintaining a frozen backbone (Huang et al., 2024). Token-level prompt tuning has been applied directly to driving benchmarks. One source-free domain adaptation study demonstrated that replacing backbone updates with visual prompt tuning enables successful adaptation from synthetic (GTA5/SYNTHIA) to real-world (Cityscapes) domains, revealing that per-layer prompt tokens can encapsulate the majority of the transfer signal even when the core network gradients are inaccessible (Ma et al., 2023). A follow-up work targeting adverse driving scenes proposes severity-aware prompt tuning that conditions prompts on weather intensity and employs an instructive chain-of-domain schedule; this approach improves semantic segmentation across multiple adverse domains without modifying the backbone weights (Gong et al., 2024b). Bird's-eye-view (BEV) perception introduces unique adaptation challenges, as it inherently couples multi-camera geometry with semantic understanding. Xiao et al. (2026; 2025a) In few-shot BEV learning, visual prompts are employed to "warm-start" prediction heads from limited data, thereby reducing the dependency on extensively labeled frames for inferring road layouts and object cues (He et al., 2024). Cross-modal alignment methods also incorporate prompt mechanisms, such as shared prompts to link camera features with language supervision for BEV retrieval and segmentation (Xie et al., 2025b). The application of prompting has also extended to end-to-end systems. For camera-only driving, injecting learned tokens into multi-modal blocks helps stabilize the policy across varied scenes and weather conditions, incurring minimal parameter overhead and requiring no changes to the backbone encoders (Liao et al., 2024). Beyond perception and control, perspective-to-BEV instruction generation utilizes prompt tuning to condition a large model on urban context (*e.g.*, lanes, signs, topology) to produce navigation instructions; here, the core challenge is to formulate prompts that ensure consistency between the model's BEV priors and the camera-view inputs (Yang et al., 2024c).

Two distinct patterns emerge from these applications. First, simple pixel-space designs remain effective when model internals are inaccessible or gradients are unavailable, and they are readily deployable in black-box settings. Second, token-level prompt tuning tends to yield an optimal trade-off between adaptation and accuracy when gradients are available, particularly with per-layer prompts or severity-aware scheduling.

Open issues include latency, as prompt generators can add overhead at scale; safety, as prompts must not trigger brittle behavior in rare, long-tail events; and calibration, since instance-adaptive prompts require uncertainty checks prior to deployment. The trajectory is evident: prompts are increasingly becoming the default mechanism for maintaining the robustness of autonomous systems as operational conditions evolve.

Ensuring operational resilience in non-stationary environments is the primary driver for PA in autonomous driving. Under these environments, weather, lighting, and sensor noise constantly fluctuate. VP serves as a compelling strategy for input-level rectification, particularly when dealing with adverse weather or when the perception module is a black box (Kalwar et al., 2023; Huang et al., 2024). In these cases, pixel-space prompts effectively act as a "preprocessing" layer that normalizes corrupted inputs before they reach the frozen backbone. Conversely, VPT is the preferred approach for addressing systemic domain shifts, such as Sim-to-Real transfer or cross-sensor BEV alignment (Ma et al., 2023; He et al., 2024). By accessing internal gradients, learnable tokens can recalibrate the deep feature space to bridge the gap between synthetic training data and real-world diversity, providing a more robust adaptation mechanism than surface-level pixel manipulation could achieve alone.

Table 2: **Comprehensive summary of VP methods across domains.**

Domain	Method	Task	Sub-type
Medical	CusSAM (Zhang & Liu, 2023)	Spatial prompts for medical segmentation	VP-Learned
Medical	SAM-Ada (Gong et al., 2024a)	3D volumetric prompting for radiology	VP-Learned
Medical	Ma-SAM (Chen et al., 2024a)	Modality-agnostic prompt encoder for multimodal	VP-Learned
Medical	FVP (Wang et al., 2023c)	Frequency prompting for cross-domain generalization	VP-Learned
Medical	DDFP (Yin et al., 2025)	Data-dependent prompting for domain shift	VP-Learned
Medical	PASS (Zhang et al., 2024a)	Test-time prompting for style/shape adaptation	VP-Learned
RS	RSPrompter (Chen et al., 2024b)	Instance segmentation prompts in remote sensing	VP-Learned
RS	IA Instance (Li et al., 2025f)	Generalizable zero-shot instance segmentation	VP-Learned
RS	ZoRI (Huang et al., 2025a)	Discriminative zero-shot RS recognition	VP-Learned
RS	PHTrack (Chen et al., 2024c)	Spectra prompting for hyperspectral tracking	VP-Learned
RS	SPTrack (Guo et al., 2024)	Spectral prompts for tracking and classification	VP-Learned
Industrial	Unsupervised AM (Era et al., 2023)	Promptable segmentation in additive manufacturing	VP-Fixed
Industrial	SAA+ (Cao et al., 2023)	Training-free prompt for anomaly segmentation	VP-Fixed
Industrial	SPT (Yang et al., 2025)	Self-perception tuning of SAM masks at test time	VP-Learned
Industrial	SAID (Huang et al., 2025b)	Scene prompts for industrial defect segmentation	VP-Learned
Industrial	ClipSAM (Li et al., 2025e)	CLIP-guided semantic/spatial prompts for SAM	VP-Generated
Industrial	CLIP→SAM (Hou et al., 2024)	CLIP coarse localization + SAM refinement	VP-Generated
Industrial	IAPAS (Zhang et al., 2025c)	Image-adaptive per-sample prompt generation	VP-Generated
Autonomous	DiffPrompter (Kalwar et al., 2023)	Differentiable implicit prompts for road segmentation	VP-Learned
Autonomous	SAMDA (Wang et al., 2024d)	Semantic prompts for adverse-weather datasets	VP-Fixed
Autonomous	SSPrompt (Huang et al., 2024)	Learnable pixel prompts replacing fixed prompts	VP-Learned
3D	P2P (Wang et al., 2022d)	Point-to-pixel prompting for 3D transfer	VP-Learned
Underwater	WaterSAM (Hong et al., 2024)	LoRA + underwater segmentation prompts	VP-Learned
Underwater	USIS-SAM (Lian et al., 2024)	Saliency prompt generation for underwater imagery	VP-Generated
Underwater	UWSAM (Li et al., 2025c)	End-to-end underwater prompt generator	VP-Generated
Adverse	SAM-EDA (Wang et al., 2024d)	Semantic prompts for weather segmentation	VP-Learned

4.6 3D Point Clouds and LiDAR

Data from 3D sensors like LiDAR, often represented as point clouds, exhibits a sparse, irregular, and permutation-invariant structure that fundamentally differs from the dense grid of 2D images. Despite these structural disparities, PA has proven transferable to this domain with only minor modifications to the model interface. Within this domain, two primary patterns have emerged: (i) input-space prompting, which reformats 3D signals for consumption by a frozen 2D or 3D encoder, and (ii) token-level prompting, which

inserts learnable tokens directly into point cloud Transformers while the backbone remains fixed. One line of research focuses on adapting 2D pretrained models to 3D tasks by *prompting the input*. Point-to-pixel prompting, for instance, converts a point cloud into geometry-preserving renderings, enabling a frozen image model to be tuned for 3D tasks using only lightweight prompt parameters (Wang et al., 2022d). Related efforts transfer CLIP to point clouds via image–depth pretraining or multi-view rendering, subsequently attaching small adapters or prompts to facilitate few- or zero-shot recognition (Huang et al., 2023c; Zhu et al., 2023). These methods demonstrate that input-space prompting can effectively repurpose large 2D or vision–language encoders for point-cloud classification, segmentation, and detection without full fine-tuning. The second, and now mainstream, direction is *token-level prompting* for 3D Transformers. Instance-aware Dynamic Prompt Tuning (IDPT) generates sample-specific prompt tokens that account for point-level noise and shape variation, improving transfer over static prompts with a minimal parameter budget (Zha et al., 2023). Dynamic Adapter Meets Prompt Tuning (DAPT) further couples per-token adapters with internal prompt tokens while freezing the backbone, yielding strong accuracy–efficiency trade-offs on benchmarks like ScanObjectNN, ModelNet40, and ShapeNetPart. Beyond generic prompts, more sophisticated designs inject geometry-aware priors, such as surface normals or curvature, directly into the prompt stream; GAPrompt, for example, introduces a point-wise prompt branch and a propagation mechanism to encode global and local geometry while maintaining a low count of trainable parameters (Ai et al., 2025). A complementary approach treats the positional encodings themselves as a form of prompt: positional prompt tuning revisits 3D positional codes to learn compact position prompts that efficiently aggregate multi-scale structure (Zhang et al., 2024c). To ensure stability across domains, Point-PRC regularizes the interaction between task-specific prompts and task-agnostic knowledge in large 3D backbones, thereby improving domain generalization without modifying frozen weights (Sun et al., 2024a). A practical strategy for maximizing parameter efficiency involves pairing prompts with low-rank adapters; PointLoRA demonstrates how LoRA modules, combined with token selection, can reduce trainable parameters while preserving performance on point cloud Transformers (Wang et al., 2025a). Prompting has also proven beneficial for multi-sensor 3D perception. In camera–LiDAR fusion, lightweight prompters can inject LiDAR-aware cues into camera-based 3D detectors to enhance depth reasoning. This technique adds negligible computational overhead and can be deployed at test time, even in the absence of LiDAR data (Guo & Ling, 2024). Modern visual–LiDAR 3D detection frameworks now employ soft prompts, inserted at various fusion stages, to guide cross-modal attention, an approach shown to achieve greater data efficiency on benchmarks like nuScenes (Li et al., 2025d). Prompt-based designs are also being explored for open-world retrieval and recognition tasks involving 3D queries, where negative prompts and complementary prompt heads have been shown to improve robustness against distribution shifts and unseen categories (Xu et al., 2024b).

Across these diverse studies, a clear pattern emerges: prompts serve as a lightweight yet controllable interface to steer large 2D or 3D backbones toward specialized point-cloud tasks. Input-space prompts are ideal for scenarios that involve repurposing existing image-based pretraining or require a model-agnostic solution. Conversely, token-level prompts are better suited for native 3D Transformers, where they can be injected as a few learnable tokens on a per-layer or per-instance basis. Geometry-aware and instance-adaptive variants achieve performance nearly on par with full fine-tuning while substantially reducing memory requirements. Finally, fusion-oriented prompting enables the seamless integration of multi-modal data, such as LiDAR and RGB, without necessitating backbone retraining.

The application of PA in 3D vision reveals a clear dichotomy based on the architecture of the pre-trained backbone. VP, often implemented via multi-view rendering or point-to-pixel projection, serves as a crucial modality bridge (Wang et al., 2022d; Zhu et al., 2023). Its primary utility lies in formatting sparse, irregular point clouds into dense, image-like inputs, enabling the direct reuse of powerful 2D vision-language models (*e.g.*, CLIP) without the need for expensive 3D pre-training. In contrast, VPT is the method of choice for adapting native 3D Transformers (Zha et al., 2023; Ai et al., 2025). Because 3D encoders must process permutation-invariant sets, injecting learnable tokens into the self-attention mechanism allows for the precise capture of local geometric structures and surface normals while maintaining high parameter efficiency (which is often lost in projection-based VP approaches).

4.7 Video Understanding and Temporal Perception

Video understanding tasks introduce a temporal dimension that complicates standard recognition, presenting challenges such as motion blur, frame-rate variability, and the need to model long-range dependencies. PA offers a lightweight mechanism for integrating temporal reasoning into frozen backbones, thereby obviating the need for full fine-tuning. A primary line of research involves augmenting video Transformers with *spatio–temporal prompt tokens*. Space–Time Prompting, for example, inserts prompts at selected layers and learns them end-to-end for class-incremental action recognition, which circumvents catastrophic forgetting while keeping the backbone frozen (Pei et al., 2023). Attention Prompt Tuning steers the self-attention mechanism using a small set of learnable prompts to improve action recognition with modest computational overhead (Bandara & Patel, 2024). To enhance feature stability against noise and camera motion, the Spatio-Temporal Prompting Network employs prompts that are shared across frames within a clip, thereby improving the robustness of the extracted video features (Sun et al., 2024c). Another practical approach operates within the compressed domain: Compressed Video Prompt Tuning learns prompts directly from motion vectors and residuals, a method that reduces decoding and training costs while maintaining competitive accuracy (Li et al., 2023a; 2025b).

Prompting is also utilized to adapt pretrained image–text models for video-centric tasks. For example, *Vita-CLIP* learns prompts for both the vision and text encoders of CLIP, enabling a single frozen backbone to balance performance between supervised and zero-shot action recognition (Wasim et al., 2023). Similarly, *TC-CLIP* injects temporally contextualized prompts that summarize clip dynamics, thereby producing stronger video-language alignment for retrieval and recognition tasks (Wang et al., 2024c). Beyond static prompts, several methods dynamically generate or update prompts on a per-clip basis. STOP integrates spatial–temporal *dynamic* prompting, wherein a lightweight module predicts prompts conditioned on the input video, improving open-domain video understanding with a frozen backbone (Liu et al., 2025d). TP-CLIP investigates whether simple *temporal prompting* is sufficient for CLIP-based video recognition, demonstrating that clip-level temporal prompts can close a significant portion of the performance gap compared to more parameter-heavy adapters (Gowda et al., 2025). Finally, PA methods have also proven effective in test-time and low-label regimes. *DTS-TPT* performs test-time prompt tuning with dual temporal synchronization to adapt to distribution shifts without requiring labels or gradient access to the backbone (Yan et al., 2024). Subsequent work extends this concept by creating a small support set dynamically and updating prompts at inference time to facilitate zero-shot video classification (Yan et al., 2025). Collectively, these results suggest an effective strategy: augmenting a frozen encoder with a few learned or generated spatio-temporal prompts that are specifically designed to target the temporal cues absent in the original model. This approach retains computational efficiency while addressing the key failure modes inherent to adapting static image models for video.

Adapting vision models to video tasks fundamentally requires bridging the gap between static spatial representation and dynamic temporal reasoning. In this context, VPT has established itself as the standard paradigm (Pei et al., 2023; Wang et al., 2024c). Since the most powerful vision backbones are pre-trained on static images (*e.g.*, ImageNet), they inherently lack temporal awareness. Injecting learnable prompt tokens—specifically designed to encode time or motion—provides a parameter-efficient way to expand the model's attention mechanism from 2D spatial correlation to 3D spatio-temporal modeling, effectively "upgrading" the architecture without the massive cost of video pre-training. While VP appears in specialized niches like compressed-domain adaptation (Li et al., 2023a), token-level injection remains the most effective method for capturing long-range temporal dependencies and action dynamics.

4.8 Underwater and Adverse Environments

Underwater scenes and adverse weather conditions (*e.g.*, fog, rain, snow, low light) introduce severe domain shifts that challenge models pretrained on standard, clear-weather datasets. The underlying physics of these environments is the primary cause of degradation. In underwater settings, wavelength-dependent light attenuation and scattering cause color distortion (typically a blue-green shift) and reduce contrast, obscuring object boundaries and fine textures. Similarly, adverse weather introduces complex, non-linear corruptions: fog and haze reduce global contrast, rain produces reflective streaks and occlusions, and snow can drastically

alter scene geometry and color distributions. In response, Prompt-based Adaptation (PA) has emerged as a powerful and practical methodology for enhancing model robustness in these specialized domains, often without requiring costly full-model retraining.

One prominent strategy involves adapting large-scale, promptable segmentation models like SAM through pixel-space visual prompts (VP). This approach is particularly effective because SAM's architecture is inherently designed to condition its output on spatial cues. By injecting prompts and fine-tuning lightweight adapters, these methods steer the frozen backbone toward domain-specific features. *WaterSAM*, for example, attaches low-rank adapters (LoRA) to SAM and uses traditional box or point prompts to segment organisms and objects in challenging underwater imagery (Hong et al., 2024). Taking this further, *USIS-SAM* develops a salient-feature prompter that automatically generates instance-specific cues, proving that intelligent prompt design significantly enhances recall in turbid water (Lian et al., 2024). A more recent pipeline, *UWSAM*, integrates an end-to-end underwater prompt generator to automatically synthesize prompts for diverse categories, demonstrating strong performance on its curated UIIS10K dataset (Li et al., 2025c). This paradigm extends effectively to terrestrial challenges as well. For instance, *SAM-EDA* uses semantic prompts to guide SAM for road segmentation in adverse weather, employing a teacher–assistant distillation scheme to transfer the performance gains into a compact student model (Wang et al., 2024d). The key advantage of these systems is their modularity; they preserve the fixed model internals and are compatible across different backbones, a crucial feature when only an API or a frozen checkpoint is accessible. When full model gradients are available, injecting learnable token prompts directly into the network (VPT) offers a more potent mechanism for managing distribution shifts. These token prompts act as learnable "instructions" prepended to the input sequence of a Vision Transformer, conditioning the self-attention layers to focus on robust, domain-invariant features while ignoring nuisance variables. *DiffPrompter* introduces differentiable implicit prompts that are learned end-to-end for road segmentation, demonstrating resilience in adverse weather (Kalwar et al., 2023). Similarly, *CoDA* proposes a severity-aware visual prompt tuning mechanism for unsupervised domain adaptation. It groups scenes by difficulty (*e.g.*, light vs. heavy fog) and trains distinct prompt branches for low- and high-severity images. Although the prompts are discarded at inference, their influence is baked into the adapted model parameters, improving its generalization (Gong et al., 2024b). For underwater semantic segmentation, *SEA-Net* applies VPT with severity perception and cross-domain priors to enhance performance across different sites and turbidity levels (He et al., 2025b). Orthogonal to severity cues, frequency-space information can be integrated with token prompts. *VFPT*, for example, uses Fourier components to guide features toward stable spectra, thereby improving robustness against weather-related corruptions with a negligible parameter increase (Zeng et al., 2024b). In summary, both underwater and adverse weather applications demand models with (i) strong geometry and shape priors to reconstruct boundaries degraded by scattering and noise, and (ii) explicit mechanisms to control for environmental nuisance factors like turbidity, haze, and illumination. Pixel-space prompting (VP) offers a straightforward, black-box-friendly solution that pairs effectively with SAM-style interfaces and PEFT techniques like LoRA. In contrast, token-level VPT affords finer-grained, white-box control over the model's internal representations when end-to-end training is feasible. A promising future direction lies in hybrid approaches that combine the strengths of both: an instance-adaptive generator could provide initial spatial prompts (VP), while a set of learned, condition-aware tokens (VPT) could simultaneously steer the model's feature extraction process to counteract the specific type and severity of environmental degradation.

Navigating the optical challenges of underwater and adverse weather environments requires addressing two distinct forms of degradation: spatial ambiguity and feature distortion. VP excels at the former, serving as a spatial anchor (Hong et al., 2024; Lian et al., 2024). In turbid or hazy conditions where object boundaries dissolve, pixel-level prompts (whether manual points or generated masks) provide the necessary localization cues to guide segmentation models like SAM, effectively compensating for the loss of high-frequency edge information. VPT, however, addresses the latter by enforcing feature invariance (Kalwar et al., 2023; Gong et al., 2024b). By tuning internal tokens, these methods recalibrate the backbone's attention to suppress environmental nuisance factors (*e.g.*, the "blue-green" shift of water or the white noise of fog), ensuring that semantic representations remain robust even when the global distribution shifts drastically from the clear-weather pre-training domain.

Table 3: **Comprehensive summary of VPT methods across domains.**

Domain	Method	Task	Sub-type
Medical	ProSFDA (Hu et al., 2022)	Source-free domain adaptation with prompt tuning	VPT-Learnable
Medical	Biomed-DPT (Peng et al., 2025)	Dual-modality visual prompt for medical reasoning	VPT-Learnable
Medical	FedLPPA (Lin et al., 2024a)	Federated personalized prompt for segmentation	VPT-Learnable
RS	De-Prompting (Liu et al., 2023a)	Temporal prompt tuning for change captioning	VPT-Learnable
RS	VPT-CLIP (Liu et al., 2025c)	Bi-temporal change detection via CLIP priors	VPT-Learnable
RS	PEFTT (Yuan et al., 2023)	Prompt-based lightweight satellite–caption retrieval	VPT-Learnable
RS	RLita (Zhang et al., 2025b)	Prompt-tuned cross-modal retrieval in RS	VPT-Learnable
RS	MVP (Zhu et al., 2024a)	Few-shot classification with spatial prior prompts	VPT-Learnable
RS	UPETU (Dong et al., 2024b)	Universal PEFT encoder for RS backbones	VPT-Learnable
RS	LayerLink (Zhu et al., 2025)	Layer-wise linking for efficient prompt adaptation	VPT-Learnable
3D	PointCLIP V2 (Zhu et al., 2023)	CLIP-to-point cloud alignment via token prompts	VPT-Learnable
3D	CLIP2Point (Huang et al., 2023c)	Image–point cloud alignment using token prompts	VPT-Learnable
3D	IDPT (Zha et al., 2023)	Instance-aware dynamic token prompts	VPT-Generated
3D	GAPrompt (Ai et al., 2025)	Geometry prompts for local and global structure	VPT-Learnable
3D	PosPrompt3D (Zhang et al., 2024c)	Positional prompt tuning for 3D transformers	VPT-Learnable
3D	Point-PRC (Sun et al., 2024a)	Regularization between task and general prompts	VPT-Learnable
3D	PointLoRA (Wang et al., 2025a)	Low-rank adapters + prompts for 3D Transformers	VPT-Learnable
3D	PromptDet (Guo & Ling, 2024)	Camera–LiDAR fusion with soft prompts	VPT-Learnable
3D	PF3Det (Li et al., 2025d)	Multi-stage LiDAR–camera fusion with prompts	VPT-Learnable
3D	NPCP (Xu et al., 2024b)	Negative prompts for robust 3D retrieval	VPT-Learnable
Autonomous	UniUVPT (Ma et al., 2023)	Source-free domain adaptation for driving datasets	VPT-Learnable
Autonomous	CoDA (Gong et al., 2024b)	Severity-aware visual prompt for adverse driving	VPT-Learnable
Autonomous	ViProBEV (He et al., 2024)	Few-shot BEV perception using visual prompts	VPT-Learnable
Autonomous	BEVCLIP (Xie et al., 2025b)	Shared prompts for BEV retrieval/segmentation	VPT-Learnable
Autonomous	PromptE2E (Liao et al., 2024)	End-to-end driving using learned tokens	VPT-Learnable
Autonomous	BEVInstructor (Yang et al., 2024c)	Perspective-to-BEV generation via prompt	VPT-Learnable
Video	ST-Prompting (Pei et al., 2023)	Spatio-temporal token prompts for video recognition	VPT-Learnable
Video	APT (Bandara & Patel, 2024)	Attention steering with learnable prompts	VPT-Learnable
Video	STPN (Sun et al., 2024c)	Shared prompts across frames for video features	VPT-Learnable
Video	CV-PT (Li et al., 2023a)	Prompts learned from motion vectors/residuals	VPT-Learnable
Video	Vita-CLIP (Wasim et al., 2023)	Dual (vision/text) prompts for zero-shot recognition	VPT-Learnable
Video	TC-CLIP (Wang et al., 2024c)	Temporally contextualized prompts for video–text	VPT-Learnable
Video	STOP (Liu et al., 2025d)	Dynamic generator producing per-clip prompts	VPT-Generated
Video	TP-CLIP (Gowda et al., 2025)	Clip-level prompting for efficient adaptation	VPT-Learnable
Video	DTS-TPT (Yan et al., 2024)	Test-time dual-synchronization prompt tuning	VPT-Learnable
Video	TESTV (Yan et al., 2025)	Dynamic support set and prompt update at inference	VPT-Learnable
Underwater	SEA-Net (He et al., 2025b)	Severity-aware VPT for underwater segmentation	VPT-Learnable

5 PA under Practical Constraints

In the previous discussions, we focus on PA attempts on conventional supervised finetuning (§2-4). Currently, PA successfully demonstrates remarkable effectiveness in a variety of learning scenarios defined by significant operational constraints Oh et al. (2023); Zhang et al. (2025e;d); Yu et al. (2023); Khattak et al. (2023). The reason is due to its parameter-efficient and data-efficient nature, making it an ideal candidate for situations where data is scarce, data distributions are non-stationary Zhang et al. (2025e); Wang et al. (2022c); Zhao et al. (2024a), or access to model internals Oh et al. (2023); Tsai et al. (2020) and computational resources is limited Zhang et al. (2025e;d); Niu et al. (2024); Zhao et al. (2025a). In this section, we categorize and survey the application of PA across these challenging paradigms. Formally, we organize these paradigms

into three core areas: adaptation under data constraints (see §5.1), adaptation in dynamic environments (see §5.2), and adaptation with resource or access limitations (see §5.3).

5.1 Data-Constrained Adaptation

Data-constrained adaptation includes scenarios where the primary limitation is the quantity or quality of labeled data. Under this condition, the proposed methods should maximize their learning effectiveness from minimal supervision or even entirely unlabeled datasets.

5.1.1 Few-Shot Learning

In the few-shot learning (FSL) setting, a model must generalize from a very small number of labeled examples. Fully fine-tuning a large-scale model in FSL risks severe overfitting, while VPT can potentially preserve the robust, generalizable features of the frozen backbone. A key challenge is the "Base-New Trade-off," where optimizing for seen (base) classes degrades performance on unseen (new) classes. Advanced methods address this by creating more dynamic and context-aware prompts. For instance, *MaPLe* (Khattak et al., 2023) introduces multi-modal prompt learning, creating learnable prompts in both the vision and language encoders of a Vision-Language Model (VLM). Crucially, it uses a "coupling function" to link the vision and language prompts, ensuring they are optimized synergistically to improve cross-modal alignment and enhance generalization from limited data.

5.1.2 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) aims to adapt a model trained on a labeled source domain to an unlabeled target domain. While many modern techniques are presented in the context of Test-Time Adaptation (TTA), their core mechanisms are directly applicable to the offline UDA setting. In this context, prompts must be optimized without direct supervision. For example, the *DePT* (Gao et al., 2022) tunes only source-initialized prompts at test time, using online pseudo-labeling and a hierarchical self-supervised regularization to adapt efficiently, even with very limited target data. Similarly, *OT-VP* (Zhang et al., 2025d) provides a principled solution by framing UDA as a distribution alignment problem. It learns a universal visual prompt for the target domain by minimizing the Optimal Transport (OT) distance between the feature distributions of the prompted target data and the source data. These approaches highlight how VPT can effectively bridge the domain gap in unsupervised settings.

5.1.3 Multimodal Learning with Missing Modalities

A significant real-world challenge is handling multimodal data where one or more modalities (*e.g.*, text accompanying an image) may be missing during training or inference. PA offers an effective, parameter-efficient solution. (Lee et al., 2023) introduce modality-missing-aware prompts, where different learnable prompts are assigned to different missing-modality cases (*e.g.*, image-only, text-only, complete). This allows a frozen multimodal transformer to adapt its behavior based on the available modalities. Building on this, *DPL* (Lu et al., 2025) proposes a decoupled prototype-based output head that can be integrated with prompt-based methods. This work uses missing-case-aware, class-wise prototypes for each modality, further improving the model's robustness by specializing the classification head itself to the missing modality scenario.

5.2 Dynamic Adaptation

Dynamic adaptation refers to scenarios in which the data distribution evolves over time, necessitating continuous and efficient model adjustment.

5.2.1 Test-Time Adaptation

Test-Time Adaptation (TTA) involves adapting a model to a new target domain during inference, often in an online setting where data arrives sequentially. The efficiency of VPT is highly suitable for this

paradigm. *DynaPrompt* (Xiao et al., 2025f) introduces dynamic test-time prompt tuning, which generates input-dependent prompts for more precise adaptation. For scenarios where the distribution continuously shifts, *DPCore* (Zhang et al., 2025e) maintains a coreset of dynamic prompts to efficiently adapt to evolving domains without catastrophic forgetting. Furthermore, some methods operate under even stricter constraints; *FOA* (Niu et al., 2024) developed a TTA method requiring only forward passes, making it extremely efficient. In the zero-shot setting, *PromptAlign* (Samadh et al., 2023) uses distribution alignment to adapt prompts at test time for generalization to unseen classes and domains.

5.2.2 Continual and Incremental Learning

Continual Learning (CL) involves training on a sequence of tasks without forgetting prior ones. VPT-based approaches reframe this from a weight-regularization problem to a prompt management problem. The seminal work *L2P* (Wang et al., 2022c) introduces a "prompt pool" that acts as a key-value memory. For any input, the model queries the pool to select the most relevant prompts to prepend to the input sequence, allowing it to dynamically compose "instructions" for the frozen backbone. Building on this, *DualPrompt* (Wang et al., 2022b) refines the approach by drawing inspiration from Complementary Learning Systems (CLS) theory. Instead of a single pool, it explicitly decouples knowledge into two sets of prompts: a shared G-Prompt (General) to learn task-invariant knowledge and a set of E-Prompts (Expert) to capture task-specific knowledge. This design more effectively isolates knowledge for dissimilar tasks while sharing common knowledge, leading to further reductions in catastrophic forgetting.

5.3 Resource and Access-Constrained Adaptation

Resource and access-constrained adaptation focuses on limitations imposed by model accessibility and the computational environment, such as black-box APIs and decentralized data settings.

5.3.1 Black-Box and Gray-Box Model Adaptation

In a black-box setting, a model is only accessible via an API, with no access to its internal parameters or gradients, making standard prompt tuning impossible. To solve this, methods have turned to gradient-free optimization. *BlackVIP* (Oh et al., 2023) pioneers this by using a small generator network (*i.e.*, a "Coordinator") to create an input-dependent visual prompt. The parameters of this tiny generator are then optimized using a zeroth-order algorithm that estimates gradients by making a small number of queries to the black-box model. This builds on earlier ideas of model reprogramming, which showed that black-box models could be repurposed for new tasks with scarce data (Tsai et al., 2020). For Vision-Language Models offered as a service, other black-box prompt tuning methods have been developed to optimize prompts in a derivative-free manner (Yu et al., 2023).

5.3.2 Federated and Decentralized Learning

In Federated Learning (FL), data is distributed across clients and cannot be centralized, posing challenges of communication overhead and statistical heterogeneity (*i.e.*, non-IID data). VPT is a natural fit, as it drastically reduces communication costs by requiring clients to transmit only small prompt parameters. To handle non-IID data, personalized FL approaches are emerging, as demonstrated in works like *pFedPrompt* (Guo et al., 2023) and *FedPrompt* (Zhao et al., 2023). In this paradigm, clients learn a combination of shared global prompts, which are aggregated at a server to capture collective knowledge, and private local prompts, which are trained only on the client's data to capture its unique distribution. This concept of decoupling prompts, central to *pFedPrompt*, allows for personalized models that benefit from collaboration while being tailored to local data characteristics.

6 Trustworthy AI

PA is increasingly used as a lightweight alternative to full fine-tuning. In trustworthy AI, PA mainly supports three goals: robustness (see §6.1), fairness and bias mitigation (see §6.2), and privacy and security (see §6.3). Below, we review representative designs and summarize practical takeaways for deployment.

6.1 Robustness

Robustness is a core requirement for trustworthy AI: models should be reliable under distribution shifts and resist adversarial manipulations. Large pre-trained models still degrade in these cases. PA on pixels (VP) or internal tokens (VPT) offers a low-cost handle to improve robustness without touching the backbone. We discuss two threads: *domain shift* and *adversarial robustness*.

Domain shift. VPT is a natural tool for cross-domain transfer. According to target access, methods fall into domain generalization (DG; no target data) and unsupervised domain adaptation (UDA; unlabeled target seen in training).

For DG, (Li et al., 2022a) introduce common–specific prompts to capture domain-shared and sample-specific cues. (Bai et al., 2024b) propose a soft prompt generator that emits instance-specific prompts conditioned on domain information. *EPVT* (Yan et al., 2023) uses a low-rank prompt generator to reduce artifact bias in lesion recognition. *DAPSAM* (Wei et al., 2024) adapts SAM via domain-adaptive prompts from prototype memories, improving cross-domain performance. On the VL side but with **visual** prompts, (Cheng et al., 2024) disentangle text-guided visual prompts into domain-invariant and domain-specific parts; *ODG-CLIP* (Singha et al., 2024) attaches specialized prompts for unknown classes. *StyLIP* (Bose et al., 2024) separates style and content across scales; (Gupta et al., 2025) use cross-attentive prompts to mix domain- and class-generic tokens.

For UDA, *USDAP* (Shao et al., 2024) learns prompts to align target and source distributions. (Zhan et al., 2024) improves robustness to noisy pseudo labels with dynamic mask prompting. In adversarial UDA, (Jin et al., 2023) meta-optimizes prompts to relabel target samples adaptively; *ADAPT* (Cui et al., 2025) employs visual prompts in a minimax game to align both global and category distributions. In medical segmentation, *ProSFDA* (Hu et al., 2022) and *DDFP* (Yin et al., 2025) reduce cross-domain gaps with prompt-based styles or frequency cues.

Adversarial robustness. Beyond natural shift, attacks craft imperceptible perturbations to force errors. Adversarial training is strong but costly; PA gives a cheaper alternative.

APD (Luo et al., 2024) learns prompts that align perturbed features with robust embeddings, improving the robustness–accuracy trade-off. *AMPT* (Zhao et al., 2025b) uses a pool of prompts and combines them dynamically. (Zhou et al., 2024c) improves few-shot robustness by enforcing consistency between clean and adversarial features. *ER-APT* (Jia et al., 2025) optimizes region-level prompts via evolutionary search to resist diverse attacks.

Test-time and structural prompting are also explored. *RobustMAE* (Huang et al., 2023a) inserts frequency-domain prompts to occupy high-frequency bands at inference. *PBL* (Li et al., 2023b) transfers robustness from a robust source by loosening decision boundaries. *ARVP* (Liu & Li, 2024) applies adversarial reprogramming with visual prompts in class-incremental learning to reduce forgetting. Depth-wise adversarial prompt tuning *(MDAPT)* (Li et al., 2025a) injects prompts across layers; *TAPT* (Wang et al., 2025c) adapts prompts per sample to secure zero-shot inference under strong attacks.

PA methods reduce trainable state and are easy to add on top of frozen encoders. They help when labels are scarce or compute is tight. For large activation footprints, pair PA with memory-efficient training (cf. §2.4); for open-world shifts, prefer instance-adaptive or frequency-aware prompts.

6.2 Fairness and Bias Mitigation

Prompts have also been used to mitigate social biases, although most methods target vision-language models (VLMs) and focus on the text branch or cross-modal alignment rather than the PA within the vision backbone that is the focus of this survey. Representative approaches fall into three main categories.

(1) Debiasing via Subspace Projection. *Biased Prompts* (Chuang et al., 2023) define a bias direction in the embedding space using a set of text prompts that describe the bias. By projecting this direction out of the feature space, it improves the fairness and robustness of CLIP without requiring retraining. This approach modifies only the text embeddings but is effective for both discriminative and generative models.

(2) Adversarial or Unified Debiasing with Prompts. (Berg et al., 2022) reduce gender and skin-tone biases in CLIP by prepending a small number of learnable prompts to text queries and employing adversarial training. This approach requires minimal computation and no access to the original training data. More recently, *SFID* (Jung et al., 2024) introduced a unified debiasing framework that reduces multimodal biases via feature clipping and confidence patching without degrading vision–language alignment.

(3) Joint Alignment and Debiasing of Image and Text Modalities. Recent work indicates that modifying only the text branch can degrade cross-modal alignment. To address this, (Zhang et al., 2025a) propose a method to jointly align and debias both the image and text modalities, mitigating the trade-off between performance and fairness. Other analytical studies also suggest that the vision branch is often a primary source of bias and that fairness across client groups must be considered in federated learning settings (Weng et al., 2024).

Relation to the Scope of This Survey. The methods above are primarily examples of VLM prompting or text-side correction. They provide evidence that prompts can serve as a control plane for debiasing, but they are not direct applications of PA via pixel (VP) or token injection (VPT) into the vision backbone. We reference them in this survey to: (i) provide a contrasting perspective, and (ii) highlight that by situating learnable prompts within the vision backbone (VPT) or in the pixel space (VP), debiasing and domain generalization can be implemented as parameter-efficient, modular components.

6.3 Privacy and Security

PA is also implicated in model security, particularly concerning backdoor attacks and copyright protection. While research in this area again focuses primarily on VLMs, the methodologies offer valuable insights into the threat surface and potential defenses for PA.

(1) Backdoor Attacks during the Prompt Learning Phase. *BadCLIP* (Bai et al., 2024a) demonstrates that backdoors can be implanted into CLIP during the prompt learning stage. It jointly optimizes a learnable image trigger with a trigger-aware text prompt generator to achieve a high attack success rate while preserving accuracy on clean samples. This directly shows that learnable or generated prompts are themselves part of the attack surface. More broadly, recent work has systematically evaluated and extended backdoor attacks on LVLMs, including vulnerabilities in cross-domain and instruction-tuning scenarios (Liang et al., 2025), as well as training-data-free, test-time backdoors like *AnyDoor* (Lu et al., 2024). These findings suggest that even with a frozen backbone, adaptation that relies solely on prompts requires corresponding security audits.

(2) Prompt-based Detection and Defense. One line of defense uses prompt tuning to detect backdoored samples, for instance, by discriminating trigger consistency or separability (Stein et al., 2024). Another approach introduces mechanisms like repulsive activation at the representation level for unified defense. Although mostly validated on VLMs, these ideas can be transferred to VPT, for example by using a small discriminator head or generator to evaluate anomalous coupling between prompts and representations.

(3) Prompt Copyright and Watermarking. *WVPrompt* (Ren et al., 2024) treats a watermark as a backdoor injected into prompts, enabling remote copyright verification through statistical tests. This work highlights the security property of *prompts as assets*. For VPT systems that use generators to produce prompts, such watermarks could also serve as a mechanism for verifying usage compliance and tracking model versions.

7 Foundational Analysis and Theory of PA

Understanding the theoretical foundations of PA is essential beyond its demonstrated practical advantages. However, we should acknowledge that the related theoretical analysis remains limited in the current community. In this section, we aim to explore some of the fundamental questions, where some questions are general (*i.e.*, Q1—3), some are specified for VP or VPT (*i.e.*, Q4, Q5).

- ***General Q1: How does PA induce behavioral changes in the model?***

  For both VP and VPT paradigms, gradient-weighted class activation mapping (GradCAM) (Selvaraju et al., 2017; Chakraborty et al., 2022; Zeng et al., 2024b) offers an intuitive visual explanation of PA's decision-making process.

  For VP, (Rezaei et al., 2024) shows that learned prompts can explicitly steer the attention of vision models with GradCAM visualizations. In particular, their experiments reveal that the added prompt pixels bias the model's attention maps toward the spatial location of the prompts, thereby altering which regions of the image the model emphasizes during decision-making. This suggests that VP modifies early feature activations in a way that reallocates attention, providing a concrete mechanism by which the prompt drives behavioral change in the frozen model.

  For VPT, (Han et al., 2024) uses the attention visualizations to reveal behavioral differences between VPT and full fine-tuning. The GradCAM analysis highlights instances where full fine-tuning fails to recognize objects, whereas prompt tuning achieves correct classifications. In these successful cases, distinct visual explanation patterns emerge through heatmaps. For instance, in an image of a bicycle where full fine-tuning fails, prompt tuning can identify the bicycle by attending to its structural features.

  While the research on PA's influence on model behaviors remains limited (*i.e.*, partially due to PA's natural incompatibility with the linear representation hypothesis (Liu et al., 2025b; Wu et al., 2024c; Geiger et al., 2021) or other interpretability approaches such as concept-based analysis (Parekh et al., 2024) or causal probing (Geiger et al., 2021; 2024) , current approaches suggest that PA can effectively direct the model's attention to salient image regions, enhancing the overall performance.

- ***General Q2: What do visual prompts learn?***

  In learnable VP, (Chen et al., 2023) shows that visual prompts primarily learn to bridge the gap between the pretrained model's label space and the downstream target classes. Their analysis demonstrates that the quality of this implicit label mapping (*i.e.*, how well source and target categories align) directly governs VP performance. In other words, VP does not endow the model with new semantics but instead learns pixel patterns that re-map pretrained representations to new class labels, effectively repurposing existing knowledge through input-space adaptation.

  In learnable VPT, recent analysis of the training dynamics reveals that prompt tokens exhibit a specific learning pattern during fine-tuning. (Wang et al., 2024b) conducts an empirical study measuring the normalized mutual information (NMI) (Estévez et al., 2009) between prompt tokens and patch tokens across different transformer layers during training. Specifically, NMI is computed using sigmoid-normalized cross-attention between prompts and patch tokens, approximating the joint distribution as:

$$\mathrm{NMI}(P_i; E_i) = \frac{2 \times I(P_i; E_i)}{H(P_i) + H(E_i)}, \tag{10}$$

  where P_i and E_i represent the prompt tokens and patch tokens at the (i)-th transformer layer, respectively, I denotes the standard mutual information, and $H(\cdot)$ represents the entropy. Their empirical observation across four datasets (*i.e.*, CUB-200-2011, Caltech-101, Patch Camelyon, and Clevrcount) shows that visual prompts learn representations that increasingly align with the patch token distributions throughout the training process (*i.e.*, distribution of prompts for downstream contextualization gradually converges towards the distribution of patch tokens), suggesting that effective prompt learning involves establishing stronger correlations with image patch embeddings rather than learning completely independent representations.

- ***General Q3: How effective are PA methods across different adaptation settings?***

  The effectiveness of PA approaches varies with the nature of the adaptation setting. Recent studies reveal that while VP demonstrate notable robustness to distribution shifts, VPT's performance is more

sensitive to the relationship between source and target tasks and data distributions.

For VP, (Wu et al., 2024a) demonstrates that VP can exhibit notable robustness against distribution shift, highlighting its effectiveness in handling out-of-distribution (OOD) settings. Empirical evaluations on benchmarks such as WILDS show that VP achieves consistent gains compared to conventional prompting strategies and even outperforms strong baselines like linear probing and full fine-tuning in certain cases. This robustness extends beyond domain shifts: when tested on corruption datasets such as CIFAR100-C and CIFAR10-C, VP maintains strong performance under diverse perturbations, rivaling the results of full fine-tuned models. These findings suggest that VP's standalone design provides better resilience to data variations. This property makes VP a compelling choice for scenarios where models are expected to generalize under distribution shifts or noisy conditions.

For VPT, (Han et al., 2024) presents a comprehensive empirical study across 19 visual tasks in the VTAB-1k benchmark, identifying key conditions under which VPT yields superior performance. The authors show that the relative effectiveness of VPT *vs*. FT depends critically on two factors: the similarity of data distributions between the pretraining and downstream tasks, and the disparity in task objectives. Specifically, VPT tends to outperform FT when (1) the downstream task is substantially different from the pretraining objective (*e.g.*, classification *vs*. spatial reasoning tasks like counting or distance estimation), or (2) the data distributions between the source and target domains are closely aligned (*e.g.*, natural images in both cases). Analyses along these two axes often reveal that VPT is well-suited to three out of the four transfer learning scenarios defined by combinations of task and data similarity. When either the task objectives differ substantially from those seen during pretraining or the data distributions are closely aligned, VPT tends to provide better performance under low-resource settings. This is attributed to its ability to preserve the pretrained model's representations while introducing only a small set of task-specific parameters. In such cases, VPT offers a favorable trade-off between adaptation capacity and parameter efficiency. However, as the amount of downstream data increases, the performance advantage of VPT diminishes, and full finetuning may become preferable. These findings suggest that the choice of tuning strategy, particularly the prompt length and extent of parameter updates, should be informed by both the task formulation and the distributional relationship between the source and target domains.

- ***VP Q4: What is the best way to place the prompt in VP?***

  Prompt placement plays a critical role in how VP methods interact with pretrained vision models, and the optimal strategy differs notably between fixed and learnable VP variants.

  For fixed VP methods, primarily developed for segmentation-oriented tasks, prompts are directly plotted or overlaid on the input image. A representative example is SAM (Kirillov et al., 2023), where prompts such as points, boxes, or masks are spatially anchored to image regions to guide the model's attention toward target objects or segments. These spatially explicit prompts serve as conditioning cues rather than learnable parameters, making their placement inherently task-driven and interpretable. In such frameworks, the prompt's position corresponds directly to the semantic location of the object or region of interest, effectively functioning as spatial supervision for dense prediction tasks.

  In contrast, learnable VP methods adopt a padding-based placement strategy to preserve the integrity of visual content while optimizing prompt effectiveness. (Wu et al., 2024a) demonstrates that the most effective way to place visual prompts is by padding learnable pixels around the image rather than embedding them within it. Their design slightly shrinks the original image and surrounds it with a non-overlapping frame of prompt pixels, ensuring that the visual content remains intact while the prompt interacts with the model's spatial representations through positional embeddings. This "border prompt" approach prevents corruption of salient image regions, stabilizes optimization, and consistently outperforms additive or internal prompt placements across diverse datasets. While other studies, such as LoR-VP (Jin et al., 2025) and SVDP (Yang et al., 2024a), have explored interior or adaptive prompt placements for dense prediction and domain adaptation, the border-padding strategy remains the most reliable and empirically validated prompt placement for image classification with frozen vision backbones.

- ***VPT Q5: What is the optimal prompt length for VPT?***

  Determining the optimal prompt length in VPT is crucial for balancing model performance and computational efficiency. Empirical findings by (Kim et al., 2024b) have shown that the relationship between prompt quantity and fine-tuning accuracy is non-linear, refuting the assumption that more prompts always lead to better performance. Notably, reducing prompt length can result in minimal accuracy loss, with most performance degradation occurring at lower prompt ranges. This is theoretically supported by the low-rank characteristics of self-attention matrices in Vision Transformers. The rank increase of the self-attention matrix with added prompts follows a logarithmic trend:

  $$\text{rank}(\tilde{A}_{n+m}) - \text{rank}(\tilde{A}_n) = O(\log(m)). \tag{11}$$

  This explains the diminishing returns of adding more prompts, as initial prompts contribute more significantly to attention than later ones. From a computational standpoint, prompt length directly affects efficiency, with longer prompts introducing substantial overhead. Thus, the optimal prompt length should reflect a balance between accuracy and resource limits. Recent works (Han et al., 2023; Shang et al., 2025; Xiao et al., 2025c;d) also support the theoretical view that prompt quality does not inherently depend on prompt length. Instead, a small number of well-chosen prompts can effectively handle the downstream fine-tuning, emphasizing that informativeness and alignment might matter more than quantity. This observation aligns with language domain-related research (Zeng et al., 2025; Liu et al., 2025a).

8 Discussion and Challenges

Guideline for Selection: VP vs. VPT. A recurring theme throughout our survey is the trade-off between the flexibility of input-space prompting and the adaptation capacity of internal prompt tuning. Based on the empirical evidence reviewed across domains, we synthesize the following comparative guidelines:

- **Scenario 1: Black-box Access and Deployment Flexibility.** When model internals are inaccessible (*e.g.*, API-based foundation models) or when the deployment environment requires strict architectural invariance, VP is the superior choice (Oh et al., 2023). Pixel-level prompts serve as an external "plug-and-play" module that can be optimized via zeroth-order methods or generated by lightweight networks, effectively circumventing the need for gradient access. This makes VP particularly dominant in interactive segmentation (*e.g.*, SAM-based medical or industrial applications) (Kirillov et al., 2023; Chen et al., 2024a), where spatial cues (points/boxes) are more intuitive than semantic tokens.
- **Scenario 2: Deep Semantic Adaptation and Domain Shifts.** For tasks involving significant domain shifts (*e.g.*, natural images to remote sensing or medical reports) or requiring temporal/modality alignment (*e.g.*, video understanding), VPT generally yields better performance trade-offs (Han et al., 2024; Pei et al., 2023). Unlike VP, which is limited to modifying the input distribution, VPT injects learnable tokens into deep transformer layers, allowing for the direct recalibration of high-level semantic features. This capacity is essential for tasks where the pre-trained features are semantically misaligned with the downstream objective, such as in 3D point cloud classification (Zha et al., 2023) or long-tailed recognition (Dong et al., 2023).
- **Scenario 3: Efficiency Constraints.** While both methods are parameter-efficient, their memory footprints differ. If training memory is the primary bottleneck, Fixed VP or inference-only adapters should be prioritized as they bypass backbone backpropagation. However, if the goal is to maximize accuracy within a limited parameter budget (*i.e.*, $< 1\%$) while tolerating standard training memory costs, VPT Deep consistently outperforms shallow VP variants by leveraging the full expressivity of the deep backbone (Jia et al., 2022; Han et al., 2023).

In summary, the choice between VP and VPT should be driven by the *access level* (black-box vs. white-box) and the *adaptation depth* (spatial correction vs. semantic realignment) required by the target application.

Despite the significance of prompt-based adaptation methods, there are critical challenges to using it in practice. In the following, we identify some of them.

- **Safety Alignment.** Safety alignment is pivotal for the advancement of AI technologies. The safety of PA approaches should be woven into their development and deployment. Key elements in this process include interpretability, governance, and rigorous verification of model properties. As stated in §2, current PA approaches can be understood as a targeted intervention towards expected data distributions. Such interventions, however, can involve malicious actors, who may fine-tune models to generate or amplify harmful content, misinformation, or biased outputs (see §6.3). To address these concerns, mitigation strategies can be considered. These include robustness evaluations, continuous monitoring of model behavior, and systematic bias audits. Additionally, comprehensive documentation of models and training datasets, alongside transparent disclosure of any known biases introduced during model development, is essential (Liu et al., 2025b).

  In sum, aligning intervention directions with human values, goals, and expectations remains a critical challenge. Advancing this alignment requires sustained research efforts, particularly in detecting instances of misalignment and formulating corrective mechanisms.

- **Training Overhead and Stability.** While PA approaches offer enhanced training parameter efficiency, they present notable limitations during training. Due to the different characteristics of VP and VPT, we discuss them separately.

  For VP, certain approaches (Huang et al., 2023b), introduce task-specific or cluster-specific prompting. These approaches necessitate additional training data for clustering and increase both prompt parameter count and optimization complexity, thereby imposing substantial additional training overhead. VP also tends to exhibit instability: whether in design, location, or pattern, minor perturbations in prompt configuration can result in significant performance degradation. This sensitivity undermines its robustness and generalization across diverse tasks or datasets. Moreover, when VP is applied to robust source models, it often inherits a trade-off: adversarial robustness is maintained at the cost of noticeable declines in standard accuracy (Li et al., 2023b).

  For VPT, the first challenge lies in the training overhead. Although VPT reduces the per-iteration training time by updating a subset of the model parameters via gradient descent, the total training duration often increases significantly compared to full fine-tuning. This is primarily due to the need for extensive hyperparameter search across prompt length, learning rates, weight decay values, *etc.* A promising direction involves fixing certain hyperparameters to reduce the overall search space during training. Another key challenge arises from the instability of training outcomes under different initialization values (*i.e.*, random seeds). While VPT has demonstrated certain robustness across various initialization strategies (*e.g.*, *He* (He et al., 2015), *Truncated Normalization* (Paszke et al., 2019)), it exhibits greater variance (*i.e.*, more fluctuating results compared to full fine-tuning). This naturally necessitates additional training to achieve satisfactory results, thereby substantially diminishing the effectiveness of current efforts on reducing per-iteration training time. It is worth noticing that the training results of VPT are also influenced by different pre-trained strategies (*e.g.*, supervised objectives, self-supervised objectives: MAE (He et al., 2022) and MoCo v3 (Chen et al., 2021)).

  To sum up, reducing PA's training overhead and strengthening the robustness of training remain critical challenges. Intensified research efforts are needed, especially in developing training shortcuts, detecting, and developing strategies to rectify training stability.

- **Inference Latency.** PA often experiences increased inference latency due to the additional components supplemented to the input (*i.e.*, VP) or appended to the original vision models (*i.e.*, VPT). This add-on naturally introduces challenges related to additional memory consumption. In response, the research community has actively sought to alleviate such inference bottlenecks. Techniques to cut down on memory demands, both in terms of size and bandwidth, and to speed up inference computations have been devised. For example, pruning-based approaches have been developed to strategically remove certain less influential components from prompting. Other approaches, such as knowledge distillation, quantization, and the combination of memory-efficient fine-tuning (Simoulin et al., 2024; Choi et al., 2025), can also decrease memory usage and enhance computational throughput by lowering the precision of model weights and activations post-training.

- **Evaluation on Real-world Environments.** The selection of pre-trained models and the evaluation of PA, have predominantly relied on standardized academic benchmarks, such as VTAB-1k, FGVC, ImageNet. Such datasets, despite their significance in the evolution of AI, have limitations in their ability to reflect real-world characteristics accurately (*i.e.*, in other words, the robustness in distribution shifting). To truly confirm the capabilities and practical applicability of PA, it is imperative to assess them using data that is diverse, complex, and mirrors real-world scenarios. While this area remains unexplored in VP, some research on VPT has shown that with measurable data distribution shifting, the results can be noticeably different (Han et al., 2024), revealing similarities to prompt tuning approaches in NLP (Wang et al., 2022a; Chen et al., 2022). Other studies on test-time adaptation (Xiao & Snoek, 2024; Lee et al., 2022; Huang et al., 2021; Li et al., 2023c) and learning (Banerjee et al., 2021; Ma et al., 2025; Xie et al., 2025a; Wang et al., 2025c) have provided initial evidence supporting the effectiveness of distributional calibration/correction techniques. Building on this observation, future work should prioritize robust methods for closing this gap. Addressing this challenge has the potential to significantly advance the landscape of PA, enhancing its adaptability across diverse visual contexts, including complex and heterogeneous scenarios encountered in AI for Science applications.

9 Conclusion

In this survey, we provide the first comprehensive review of prompt-based adaptation (PA) methods in large vision models. We highlight the fundamental differences between the two mainstream adaptation approaches: visual prompting (VP) and visual prompt tuning (VPT), and further discuss their applications, foundational analysis and theories, and current challenges. Given PA methods as lightweight and effective alternatives to full-fine-tuning under certain critical conditions, such as limited data, we hope our survey would enable researchers and practitioners to leverage their potential and drive innovation further in this area.

References

Zixiang Ai, Zichen Liu, Yuanhang Lei, Zhenyu Cui, Xu Zou, and Jiahuan Zhou. GAPrompt: Geometry-aware point cloud prompt for 3D vision model. *arXiv preprint arXiv:2505.04119*, 2025. URL https://arxiv.org/abs/2505.04119.

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. URL https://arxiv.org/abs/2203.17274.

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. BadCLIP: Trigger-aware prompt learning for backdoor attacks on CLIP. 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Bai_BadCLIP_Trigger-Aware_Prompt_Learning_for_Backdoor_Attacks_on_CLIP_CVPR_2024_paper.pdf.

Shuanghao Bai, Yuedi Zhang, Wanqi Zhou, Zhirong Luan, and Badong Chen. Soft prompt generation for domain generalization. In *European Conference on Computer Vision (ECCV)*, pp. 434–450. Springer, 2024b.

Supun Bandara and Vishal M. Patel. Attention prompt tuning: Parameter-efficient adaptation of pre-trained models for spatiotemporal modeling. *arXiv preprint arXiv:2403.06978*, 2024. URL https://arxiv.org/abs/2403.06978.

Pratyay Banerjee, Tejas Gokhale, and Chitta Baral. Self-supervised test-time learning for reading comprehension. *arXiv preprint arXiv:2103.11263*, 2021. URL https://arxiv.org/abs/2103.11263.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 25005–25017, 2022.

Hanna Berg, Adrian O'Hagan, and Anders Søgaard. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, 2022. URL https://aclanthology.org/2022.aacl-main.61.pdf.

Hanbo Bi, Yingchao Feng, Wenhui Diao, Peijin Wang, Yongqiang Mao, Kun Fu, Hongqi Wang, and Xian Sun. Prompt-and-transfer: Dynamic class-aware enhancement for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. doi: 10.1109/TPAMI.2024.3412543.

Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. StyLIP: Multi-scale style-conditioned prompt learning for CLIP-based domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5542–5552, 2024. doi: 10.1109/WACV57701.2024.00557.

Yunkang Cao, Fei Chen, Yueming Li, Zhexing Wang, and Cheng Wang. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023. URL https://arxiv.org/abs/2305.10724.

Soumava Chakraborty, Utkarsh Trehan, Khawla Mallat, and Jean-Luc Dugelay. Generalizing adversarial explanations with grad-CAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 132–141, 2022. doi: 10.1109/CVPRW56347.2022.00021.

Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, and Sijia Liu. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19133–19143, 2023. doi: 10.1109/CVPR52729.2023.01838.

Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, et al. MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. *Medical Image Analysis*, 98:103310, 2024a. doi: 10.1016/j.media.2024.103310.

Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. Revisiting parameter-efficient tuning: Are we really there yet? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8545–8557. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.580.

Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024b. doi: 10.1109/TGRS.2024.3364654.

Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9640–9649, 2021. doi: 10.1109/ICCV48922.2021.00954.

Yuzeng Chen, Yuqi Tang, Xin Su, Jie Li, Yi Xiao, Jiang He, and Qiangqiang Yuan. PHTrack: Prompting for hyperspectral video tracking. volume 62, pp. 1–18, 2024c. doi: 10.1109/TGRS.2024.3461316.

Zhihao Chen and Yisen Wang. Attribute-based visual reprogramming for vision-language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23595–23604, 2024. doi: 10.1109/CVPR52729.2024.02008.

Gong Cheng, Xingxing Xie, Junwei Han, Lei Guo, and Gui-Song Xia. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3735–3756, 2020. doi: 10.1109/JSTARS.2020.3005403.

Ali Cheraghian, Zeeshan Hayder, Sameera Ramasinghe, Shafin Rahman, Javad Jafaryahya, Lars Petersson, and Mehrtash Harandi. Canonical shape projection is all you need for 3D few-shot class incremental learning. In *European Conference on Computer Vision (ECCV)*, pp. 36–53. Springer, 2024.

Daewon Choi, Jimin Lee, Jihoon Tack, Woomin Song, Saket Dingliwal, Sai Muralidhar Jayanthi, Bhavana Ganesh, Jinwoo Shin, Aram Galstyan, and Sravan Babu Bodapati. Think clearly: Improving reasoning via redundant token pruning. *arXiv preprint arXiv:2507.08806*, 2025. URL https://arxiv.org/abs/2507.08806.

Ching-Yao Chuang, Yingxiang Lu, and Christopher Ré. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. URL https://arxiv.org/abs/2302.00070.

Chaoran Cui, Ziyi Liu, Shuai Gong, Lei Zhu, Chunyun Zhang, and Hui Liu. When adversarial training meets prompt tuning: Adversarial dual prompt tuning for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 2025.

Rajshekhar Das, Yonatan Dukler, Avinash Ravichandran, and Ashwin Swaminathan. Learning expressive prompting with residuals for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3366–3377, 2023. doi: 10.1109/CVPR52688.2022.00336.

Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. LPT: Long-tailed prompt tuning for image classification. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=P51g2v7dER.

Shaohua Dong, Yunhe Feng, Qing Yang, Yan Huang, Dongfang Liu, and Heng Fan. Efficient multimodal semantic segmentation via dual-prompt learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 14196–14203, 2024a. doi: 10.1109/IROS55529.2024.10783350.

Zhe Dong, Yanfeng Gu, and Tianzhu Liu. UPETU: A unified parameter-efficient fine-tuning framework for remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024b. doi: 10.1109/TGRS.2024.3367746.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Israt Zarin Era, Imtiaz Ahmed, Zhichao Liu, and Srinjoy Das. An unsupervised approach towards promptable defect segmentation in laser-based additive manufacturing by segment anything. *arXiv preprint arXiv:2312.04063*, 2023. URL https://arxiv.org/abs/2312.04063.

Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(11):1892–1901, 2009. doi: 10.1109/TNN.2009.2030940.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation with textual inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1262–1271, 2023. doi: 10.1109/CVPR52729.2023.00128.

Haoyue Gao et al. Prompt-based ingredient-oriented all-in-one image restoration. *arXiv preprint arXiv:2309.03063*, 2023. URL https://arxiv.org/abs/2309.03063.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. doi: 10.1007/s11263-023-01891-x.

Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N. Metaxas. Visual prompt tuning for test-time domain adaptation. 2022. URL https://arxiv.org/abs/2210.04831.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 10467–10479, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2024.

Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3DSAM-Adapter: Holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation. *Medical Image Analysis*, 98:103324, 2024a. doi: 10.1016/j.media.2024.103324.

Ziqi Gong, Hang Guo, Shuyang Sun, Hang Zhao, and Song Bai. CoDA: Instructive chain-of-domain adaptation for adverse driving scenes. In *European Conference on Computer Vision (ECCV)*, 2024b.

Raghu R. Gowda et al. TP-CLIP: Is temporal prompting all we need for temporal-aware CLIP-based video understanding? *arXiv preprint arXiv:2504.01890*, 2025. URL https://arxiv.org/abs/2504.01890.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Jing Gu, Yuxuan Li, Chen Chen, Ke Xu, and Yuxuan Wang. Prompt engineering for vision-language models: A survey. *arXiv preprint arXiv:2307.12980*, 2023. URL https://arxiv.org/abs/2307.12980.

Gaowei Guo, Zhaoxu Li, Wei An, Yingqian Wang, Xu He, Yihang Luo, Qiang Ling, Miao Li, and Zaiping Lin. SPTrack: Spectral similarity prompt learning for hyperspectral object tracking. *Remote Sensing*, 16 (16):2975, 2024. doi: 10.3390/rs16162975.

Kun Guo and Qiang Ling. PromptDet: A lightweight 3d object detection framework with lidar prompts. *arXiv preprint arXiv:2412.12460*, 2024. URL https://arxiv.org/abs/2412.12460.

Tao Guo, Song Guo, and Junxiao Wang. pFedPrompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference (WWW)*, pp. 1364–1374, 2023. doi: 10.1145/3543507.3583321.

Divyam Gupta, Mainak Singha, Sai Bhargav Rongali, Ankit Jha, Muhammad Haris Khan, Biplab Banerjee, et al. OSLoPrompt: Bridging low-supervision challenges and open-set domain generalization in CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, and Dongfang Liu. E2VPT: An effective and efficient approach for visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15606–15615, 2023. doi: 10.1109/ICCV51070.2023.01428.

Cheng Han, Qifan Wang, Yiming Cui, Wenguan Wang, Lifu Huang, Siyuan Qi, and Dongfang Liu. Facing the elephant in the room: Visual prompt tuning or full finetuning? In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=3Y0d6255kM.

Along He, Yanlin Wu, Zhihong Wang, Tao Li, and Huazhu Fu. DVPT: Dynamic visual prompt tuning of large pre-trained models for medical image analysis. *Neural Networks*, 185:107168, 2025a. doi: https://doi.org/10.1016/j.neunet.2025.107168.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553.

Qibin He. Prompting multi-modal image segmentation with semantic grouping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2094–2102, 2024. doi: 10.1609/aaai.v38i3.27945.

Weiyu He, Zhenqing Wang, Xingping Li, and Jian Zhang. SEA-Net: Severity-aware visual prompt tuning for underwater semantic segmentation. In *Computational Visual Media*. Springer, 2025b. doi: 10.1007/978-981-96-5759-4_3.

Zhenguo He, Weixiang Liu, Yinchao Xu, et al. ViPro-BEV: A task-general few-shot visual prompt learning method for BEV perception. In *Pattern Recognition and Computer Vision (PRCV)*. Springer, 2024. doi: 10.1007/978-981-97-0200-8_23.

Yang Hong, Xiaowei Zhou, Ruzhuang Hua, Qingxuan Lv, and Junyu Dong. WaterSAM: Adapting SAM for underwater object segmentation. *Journal of Marine Science and Engineering*, 12(9):1616, 2024. doi: 10.3390/jmse12091616.

Mir Rayat Imtiaz Hossain, Mennatullah Siam, Leonid Sigal, and James J Little. Visual prompting for generalized few-shot segmentation: A multi-scale approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23470–23480, 2024. doi: 10.1109/CVPR52729.2024.01989.

Yiyi Hou, Yuchen Fang, Ying Niu, Yansheng Xu, and Xi Li. Enhancing zero-shot anomaly detection: CLIP-SAM multi-stage collaboration for precise segmentation. In *Recent Trends in Intelligent Computing and Applications*, pp. 47–62. Springer, 2024. doi: 10.1007/978-981-97-8490-5_4.

Shishuai Hu, Zehui Liao, and Yong Xia. ProSFDA: Prompt learning based source-free domain adaptation for medical image segmentation. *arXiv preprint arXiv:2211.11514*, 2022. URL https://arxiv.org/abs/2211.11514.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 19665–19676, 2021.

Jiaxing Huang, Xizhou Zhu, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Learning to prompt segmentation foundation models. *arXiv preprint arXiv:2401.04651*, 2024. URL https://arxiv.org/abs/2401.04651.

Qidong Huang, Xiaoyi Dong, Dongdong Chen, Yinpeng Chen, Lu Yuan, Gang Hua, Weiming Zhang, and Nenghai Yu. Improving adversarial robustness of masked autoencoders via test-time frequency-domain prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1600–1610, 2023a. doi: 10.1109/ICCV51070.2023.00142.

Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10878–10887, 2023b. doi: 10.1109/CVPR52729.2023.01048.

Shiqi Huang, Shuting He, and Bihan Wen. ZoRI: Towards discriminative zero-shot remote sensing instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025a.

Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2Point: Transfer CLIP to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22157–22167, 2023c. doi: 10.1109/ICCV51070.2023.01991.

Yan Huang, Jian Li, Lei Chen, et al. Segment all industrial defects with scene prompts. *Sensors*, 25(16): 4929, 2025b. doi: 10.3390/s25164929.

Eugenia Iofinova, Alexandra Peste, Mark Kurtz, and Dan Alistarh. How well do sparse imagenet models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12656–12665, 2022. doi: 10.1109/CVPR52688.2022.01230.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19606–19616, 2023. doi: 10.1109/CVPR52729.2023.01882.

Joonhyun Jeong, Seyun Bae, Yeonsung Jung, Jaeryong Hwang, and Eunho Yang. Exploring the transferability of visual prompting for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26562–26572, 2025.

Yingrui Ji, Xi Xiao, Gaofei Chen, Hao Xu, Chenrui Ma, Lijing Zhu, Aokun Liang, and Jiansheng Chen. Cibr: Cross-modal information bottleneck regularization for robust clip generalization. In *International Conference on Artificial Neural Networks*, pp. 247–259. Springer Nature Switzerland, 2025.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pp. 709–727. Springer, 2022.

Xiaojun Jia, Sensen Gao, Simeng Qin, Ke Ma, Xinfeng Li, Yihao Huang, Wei Dong, Yang Liu, and Xiaochun Cao. Evolution-based region adversarial prompt learning for robustness enhancement in vision-language models. *arXiv preprint arXiv:2503.12874*, 2025. URL https://arxiv.org/abs/2503.12874.

Yunhe Jia, Zefan Li, Mengmeng Yang, Zhilin Wang, Dong Liu, Li Zhang, Jie Ren, Qian Zhang, and Qi Tian. Rate-distortion-cognition controllable versatile neural image compression. In *European Conference on Computer Vision (ECCV)*. Springer, 2024. doi: 10.1007/978-3-031-72814-3_9.

Songtao Jiang, Yan Zhang, Chenyi Zhou, Yeying Jin, Yang Feng, Jian Wu, and Zuozhu Liu. Joint visual and text prompting for improved object-centric perception with multimodal large language models. In *arXiv preprint arXiv:2404.04514*, 2024.

Can Jin, Ying Li, Mingyu Zhao, Shiyu Zhao, Zhenting Wang, Xiaoxiao He, Ligong Han, Tong Che, and Dimitris N. Metaxas. LoR-VP: Low-rank visual prompting for efficient vision model adaptation. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=5btFIv2PNb.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2607–2615, 2024. doi: 10.1609/aaai.v38i3.28003.

Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Domain prompt tuning via meta relabeling for unsupervised adversarial adaptation. *IEEE Transactions on Multimedia*, 26:8333–8347, 2023. doi: 10.1109/TMM.2023.3308332.

Hwiyeol Jung, Junyoung Kim, Seunghyun Yoon, and Sung Ju Hwang. A unified debiasing approach for vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Shoaib Ahmed Kalwar, Wenyu Zhang, Shijian Lu, Fumin Shen, and Guo-Jun Qi. Differentiable implicit visual prompts for object segmentation. *arXiv preprint arXiv:2310.04181*, 2023. URL https://arxiv.org/abs/2310.04181.

Chia-Hao Kao, Yi-Hsin Chen, Cheng Chien, Wei-Chen Chiu, and Wen-Hsiao Peng. Transformer-based image compression with variable image quality objectives. *arXiv preprint arXiv:2309.12717*, 2023. URL https://arxiv.org/abs/2309.12717.

Muhammad Uzair Khattak, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19122–19132, 2023. doi: 10.1109/CVPR52729.2023.01837.

Beomyoung Kim, Joonsang Yu, and Sung Ju Hwang. ECLIPSE: Efficient continual learning in panoptic segmentation with visual prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3346–3356, 2024a. doi: 10.1109/CVPR52729.2024.00293.

Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *Advances in Neural Information Processing Systems*, 36:36187–36207, 2023.

Youngeun Kim, Yuhang Li, Abhishek Moitra, Ruokai Yin, and Priyadarshini Panda. Do we really need a large number of visual prompts? *Neural Networks*, 181:120–125, 2024b. doi: 10.1016/j.neunet.2024.106208.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3904–3914, 2023. doi: 10.1109/ICCV51070.2023.00358.

Minh Le, Anh Nguyen, Huy Nguyen, Chau Nguyen, and Nhat Ho. Adaptive prompt: Unlocking the power of visual prompt tuning. *arXiv preprint arXiv:2501.18936*, 2025. URL https://arxiv.org/abs/2501.18936.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952, 2023. doi: 10.1109/CVPR52729.2023.01439.

Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466*, 2022. URL https://arxiv.org/abs/2210.11466.

Bing Lei, Fan Yang, Peng Yang, and Zhaoxiang Xie. Prompt learning in computer vision: A survey. *Frontiers of Information Technology & Electronic Engineering*, 25(9):1201–1221, 2024. doi: 10.1631/FITEE.2300389.

Aodi Li, Liansheng Zhuang, Shuo Fan, and Shafei Wang. Learning common and specific visual prompts for domain generalization. In *Asian Conference on Computer Vision (ACCV)*, pp. 4260–4275. Springer, 2022a.

Bing Li, Jiaxin Chen, Xiuguo Bao, and Di Huang. Compressed video prompt tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 31895–31907, 2023a.

Bingchen Li, Xin Li, Yiting Lu, Ruoyu Feng, Mengxi Guo, Shijie Zhao, Li Zhang, and Zhibo Chen. PromptCIR: Blind compressed image restoration with prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 6442–6452, 2024a. doi: 10.1109/CVPRW60753.2024.00551.

Chao Li, Yonghao Liao, Caichang Ding, and Zhiwei Ye. MDAPT: Multi-modal depth adversarial prompt tuning to enhance the adversarial robustness of visual language models. *Sensors*, 25(1):258, 2025a. doi: 10.3390/s25010258.

Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. MagicID: Hybrid preference optimization for ID-consistent and dynamic-preserved video customization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b. arXiv:2503.12689.

Hua Li, Qi Zhang, Zheng Fang, Chenglong Yang, and Kai Li. UWSAM and UIIS10K: Underwater segmentation with end-to-end underwater prompt generator. *arXiv preprint arXiv:2505.15581*, 2025c. URL https://arxiv.org/abs/2505.15581.

Kun Li et al. A prompted foundation feature assisted visual lidar 3d detection model. Technical Report TR2025-076, Mitsubishi Electric Research Laboratories (MERL), 2025d. URL https://www.merl.com/publications/docs/TR2025-076.pdf.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965, 2022b. doi: 10.1109/CVPR52688.2022.01066.

Qi Li, Liangzhi Li, Zhouqiang Jiang, and Bowen Wang. Towards robust and accurate visual prompting. *arXiv preprint arXiv:2311.10992*, 2023b. URL https://arxiv.org/abs/2311.10992.

Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. ClipSAM: CLIP and SAM collaboration for zero-shot anomaly segmentation. *Neurocomputing*, 618:129122, 2025e. doi: 10.1016/j.neucom.2024.129122.

Xinhao Li, Yuhan Zhu, and Limin Wang. ZeroI2V: Zero-cost adaptation of pre-trained transformers from image to video. In *European Conference on Computer Vision (ECCV)*, pp. 425–443. Springer, 2024b.

Xuexue Li, Wenhui Diao, Xinming Li, and Xian Sun. Insight any instance: Promptable instance segmentation for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2025f.

Yushu Li, Xun Xu, Yongyi Su, and Kui Jia. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6524–6533, 2023c. doi: 10.1109/ICCV51070.2023.00600.

Xin Lian, Xi Liu, Liwei Chen, Longbiao Li, Gang Li, Qingshan Liu, and Lorenzo Bruzzone. Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. *arXiv preprint arXiv:2406.06039*, 2024. URL https://arxiv.org/abs/2406.06039.

Siyuan Liang, Aishan Liu, Xiaojun Jia, et al. Revisiting backdoor attacks against large vision-language models from domain shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Xiwen Liang, Minzhe Niu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Visual exemplar driven task-prompting for unified perception in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9611–9621, 2023. doi: 10.1109/CVPR52729.2023.00925.

Qiyuan Liao, Kuan Wang, Hengjie Cai, et al. Prompting multi-modal tokens to enhance end-to-end autonomous driving. *arXiv preprint arXiv:2404.04869*, 2024. URL https://arxiv.org/abs/2404.04869.

Li Lin, Yixiang Liu, Jiewei Wu, Pujin Cheng, Zhiyuan Cai, Kenneth KY Wong, and Xiaoying Tang. FedLPPA: Learning personalized prompt and aggregation for federated weakly-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024a. doi: 10.1109/TMI.2024.3374229.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. In *arXiv preprint arXiv:2403.20271*, 2024b.

Yuanze Lin, Yunsheng Li, Dongdong Chen, Weijian Xu, Ronald Clark, Philip Torr, and Lu Yuan. Rethinking visual prompting for multimodal large language models with external knowledge. *arXiv preprint arXiv:2407.04681*, 2024c.

Chenyang Liu, Rui Zhao, Jianqi Chen, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023a. doi: 10.1109/TGRS.2023.3283282.

Shuping Liu and Wei Li. ARVP: Adversarial reprogramming visual prompts for class-incremental learning. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024. doi: 10.1109/IJCNN54924.2024.10642738.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19434–19445, 2023b.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19434–19445, 2023c. doi: 10.1109/CVPR52729.2023.01865.

Xiangyu Liu, Tianyang Zhang, and Yisen Wang. Sample-specific masks for visual reprogramming-based prompting. In *International Conference on Machine Learning (ICML)*, 2024.

Y. Liu, J. C. Liang, H. Fang, W. Yang, Y. Cui, X. Han, L. Huang, D. Liu, Q. Wang, and C. Han. All you need is one: Capsule prompt tuning with a single vector. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Yiyang Liu, James Chenhao Liang, Ruixiang Tang, Yugyung Lee, Majid Rabbani, Sohail Dianat, Raghuveer Rao, Lifu Huang, Dongfang Liu, Qifan Wang, et al. Re-imagining multimodal instruction tuning: A representation view. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025b. URL https://openreview.net/forum?id=LpBR82Bb2j.

Yuhao Liu, Zhiyong Zheng, and Renlong Hang. A remote sensing change detection network using visual-prompt enhanced CLIP. *IEEE Geoscience and Remote Sensing Letters*, 2025c.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00986.

Zebin Liu, Xiaodong Xu, Zhen Zhang, et al. STOP: Integrated spatial-temporal dynamic prompting for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025d.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

Jochem Loedeman, Maarten C Stol, Tengda Han, and Yuki M Asano. Prompt generation networks for input-space adaptation of frozen vision transformers. *arXiv preprint arXiv:2210.06466*, 2024. URL https://arxiv.org/abs/2210.06466.

Dongze Lu, Zechen Pu, Z M Tonmoy, et al. AnyDoor: Test-time backdoor attacks on multimodal large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Jueqing Lu, Yuanyuan Qi, Xiaohao Yang, Shujie Zhou, and Lan Du. Decoupled multimodal prototypes for visual recognition with missing modalities. *arXiv preprint arXiv:2505.08283*, 2025. URL https://arxiv.org/abs/2505.08283.

Lin Luo, Xin Wang, Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Adversarial prompt distillation for vision-language models. *arXiv preprint arXiv:2411.15244*, 2024. URL https://arxiv.org/abs/2411.15244.

Jing Ma, Hanlin Li, and Xiang Xiang. PTTA: Purifying malicious samples for test-time model adaptation. In *International Conference on Machine Learning (ICML)*, 2025.

Xinhong Ma, Yiming Wang, Hao Liu, Tianyu Guo, and Yunhe Wang. When visual prompt tuning meets source-free domain adaptive semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *European Conference on Computer Vision (ECCV)*, pp. 70–87. Springer, 2022.

Shentong Mo, Yansen Wang, Xufang Luo, and Dongsheng Li. LSPT: Long-term spatial prompt tuning for visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3985–3993, 2024. doi: 10.1609/aaai.v38i4.28189.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *International Conference on Machine Learning (ICML)*, pp. 38318–38332. PMLR, 2024.

Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. BlackVIP: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15995–16004, 2023. doi: 10.1109/CVPR52729.2023.01535.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. A concept-based explainability framework for large multimodal models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Wenxuan Pei et al. SA2VP: Spatially aligned-and-adapted visual prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Yixuan Pei, Zhiwu Qing, Shiwei Zhang, Xiang Wang, Yingya Zhang, Deli Zhao, and Xueming Qian. Space-time prompting for video class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11932–11942, 2023. doi: 10.1109/ICCV51070.2023.01089.

Wei Peng, Kang Liu, Jianchen Hu, and Meng Zhang. Biomed-DPT: Dual modality prompt tuning for biomedical vision-language models. *arXiv preprint arXiv:2505.05189*, 2025. URL https://arxiv.org/abs/2505.05189.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. PromptIR: Prompting for all-in-one blind image restoration. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. ShapeLLM: Universal 3D object understanding for embodied interaction. In *European Conference on Computer Vision (ECCV)*, pp. 214–238. Springer, 2024.

Shiyu Qin, Yi-Min Zhou, Jin-Peng Wang, Bin Chen, Bao-Yi An, Tao Dai, and Shu-Tao Xia. Progressive learning with visual prompt tuning for variable-rate image compression. In *IEEE International Conference on Image Processing (ICIP)*, pp. 1767–1773, 2024. doi: 10.1109/ICIP50875.2024.10636102.

Chunping Qiu, Xiaoyu Zhang, Xiaochong Tong, Naiyang Guan, Xiaodong Yi, Ke Yang, Junjie Zhu, and Anzhu Yu. Few-shot remote sensing image scene classification: Recent advances, new baselines, and future trends. *ISPRS Journal of Photogrammetry and Remote Sensing*, 209:368–382, 2024. doi: 10.1016/j.isprsjprs.2024.01.025.

Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. VCP-CLIP: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision (ECCV)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamala Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

R. Ravi et al. Autoprompt for mllms: Automated visual prompt generation via cross-modal optimization. In *arXiv preprint arXiv:2406.xxxxx*, 2024.

Haoran Ren, Yiding Liu, Zihan Chen, Zhaoxiang Zhang, and Jinyuan Jia. Are you copying my prompt? protecting the copyright of prompts via watermark verification. *arXiv preprint arXiv:2405.15161*, 2024. URL https://arxiv.org/abs/2405.15161.

Li Ren, Chen Chen, Liqiang Wang, and Kien Hua. DA-VPT: Semantic-guided visual prompt tuning for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Razieh Rezaei, Masoud Jalili Sabet, Jindong Gu, Daniel Rueckert, Philip Torr, and Ashkan Khakzar. Learning visual prompts for guiding the attention of vision transformers, 2024.

Chunqing Ruan and Hongjian Wang. Dynamic visual prompt tuning for parameter efficient transfer learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 293–303. Springer, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 13289–13303, 2023.

Marko Savic and Guoying Zhao. De-identification of facial videos while preserving remote physiological utility. In *British Machine Vision Conference (BMVC)*, 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Chikai Shang, Mengke Li, Yiqun Zhang, Zhen Chen, Jinlin Wu, Fangqing Gu, Yang Lu, and Yiu-ming Cheung. Iterative prompt relocation for distribution-adaptive visual prompt tuning. *arXiv preprint arXiv:2503.06901*, 2025. URL https://arxiv.org/abs/2503.06901.

Xun Shao, Mingwen Shao, Sijie Chen, and Yuanyuan Liu. USDAP: Universal source-free domain adaptation based on prompt learning. *Journal of Electronic Imaging*, 33(5):053015, 2024. doi: 10.1117/1.JEI.33.5.053015.

Antoine Simoulin, Namyong Park, Xiaoyi Liu, and Grey Yang. Memory-efficient fine-tuning of transformers via token selection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1565–1580. Association for Computational Linguistics, 2024.

Mainak Singha, Ankit Jha, Shirsha Bose, Ashwin Nair, Moloud Abdar, and Biplab Banerjee. Unknown prompt the only lacuna: Unveiling CLIP's potential for open domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13309–13319, 2024. doi: 10.1109/CVPR52729.2024.01120.

Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania, Han Zhang, Yuan Hao, Irfan Essa, and Lu Jiang. Visual prompt tuning for generative transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19840–19851, 2023. doi: 10.1109/CVPR52729.2023.01905.

Katharina Stein et al. Harnessing prompt tuning in vision-language models to detect backdoored samples. *arXiv preprint arXiv:2412.08755*, 2024. URL https://arxiv.org/abs/2412.08755.

Hongyu Sun, Qiuhong Ke, Yongcai Wang, Wang Chen, Kang Yang, Deying Li, and Jianfei Cai. Point-PRC: A prompt learning based regulation framework for generalizable point cloud analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Wei Sun, Qian Liu, and Yisen Wang. Unleashing the power of visual prompting at the pixel level. *Transactions on Machine Learning Research (TMLR)*, 2024b.

Yue Sun, Hao Chen, Yujie Zhong, Liqiang Nie, Meng Liu, and Quanzeng You. Spatio-temporal prompting network for robust video feature extraction. *arXiv preprint arXiv:2402.02574*, 2024c. URL https://arxiv.org/abs/2402.02574.

Yiwen Tang, Ray Zhang, Zoey Guo, Xianzheng Ma, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Point-PEFT: Parameter-efficient fine-tuning for 3D pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5171–5179, 2024a. doi: 10.1609/aaai.v38i6.28318.

Yiwen Tang, Ray Zhang, Jiaming Liu, Zoey Guo, Bin Zhao, Zhigang Wang, Peng Gao, Hongsheng Li, Dong Wang, and Xuelong Li. Any2Point: Empowering any-modality large models for efficient 3D understanding. In *European Conference on Computer Vision (ECCV)*, pp. 456–473. Springer, 2024b.

Shanshan Tian, Yuan Zhang, Yuhang Wang, Shuchang Zhang, Menghan Wang, Yuxin Ren, Jingwen Ren, Chao Wang, Zheng Gao, Gang Zeng, et al. SMC++: Masked learning with engineering breakthrough for semantic compression. *arXiv preprint arXiv:2403.08052*, 2024a. URL https://arxiv.org/abs/2403.08052.

Yuan Tian, Guo Lu, and Guangtao Zhai. Free-VSC: Free semantics from visual foundation models for unsupervised video semantic compression. In *European Conference on Computer Vision (ECCV)*, pp. 163–183. Springer, 2024b.

Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *International Conference on Machine Learning (ICML)*, pp. 9614–9624. PMLR, 2020.

Chaozheng Wang, Yuanhang Yang, Cuiyun Gao, Yun Peng, Hongyu Zhang, and Michael R Lyu. No more fine-tuning? an experimental evaluation of prompt tuning in code intelligence. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pp. 382–394, 2022a. doi: 10.1145/3540250.3549150.

Cong Wang, Jinshan Pan, Wei Wang, Jiangxin Dong, Mengzhu Wang, Yakun Ju, and Junyang Chen. PromptRestorer: A prompting image restoration method with degradation perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 8898–8912, 2023a.

Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10(1):574, 2023b. doi: 10.1038/s41597-023-02484-9.

Liyang Wang, Xuexue Li, Wenhui Diao, Xinming Li, and Xian Sun. Insight any instance: Promptable instance segmentation for remote sensing images. *arXiv preprint arXiv:2401.09014*, 2024a. URL https://arxiv.org/abs/2401.09014.

Song Wang, Xiaolu Liu, Lingdong Kong, Jianyun Xu, Chunyong Hu, Gongfan Fang, Wentong Li, Jianke Zhu, and Xinchao Wang. PointLoRA: Low-rank adaptation with token selection for point cloud learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Tianyang Wang, Xi Xiao, Gaofei Chen, Hanzhang Chi, Qi Zhang, Guo Cheng, and Yingrui Ji. TASAM: Terrain-and-aware segment anything model for temporal-scale remote sensing segmentation. In *Proceedings of the 37th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2025b. arXiv:2509.15795.

Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. TAPT: Test-time adversarial prompt tuning for robust inference in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Yan Wang, Jian Cheng, Yixin Chen, Shuai Shao, Lanyun Zhu, Zhenzhou Wu, Tao Liu, and Haogang Zhu. FVP: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(12):3738–3751, 2023c. doi: 10.1109/TMI.2023.3306105.

Yuzhu Wang, Lechao Cheng, Chaowei Fang, Dingwen Zhang, Manni Duan, and Meng Wang. Revisiting the power of prompt for visual tuning. *arXiv preprint arXiv:2402.02382*, 2024b. URL https://arxiv.org/abs/2402.02382.

Zhenzhi Wang, Jingyuan Lu, Zhe Li, et al. TC-CLIP: Temporally-contextualized prompting for video-text modeling. In *European Conference on Computer Vision (ECCV)*, 2024c.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. DualPrompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision (ECCV)*, pp. 126–144. Springer, 2022b.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 139–149, 2022c. doi: 10.1109/CVPR52688.2022.00021.

Ziquan Wang, Yongsheng Zhang, Zhenchao Zhang, Zhipeng Jiang, Ying Yu, and Lei Li. Exploring semantic prompts in the segment anything model for robust remote sensing semantic segmentation under adverse weather. volume 16, pp. 758, 2024d. doi: 10.3390/rs16050758.

Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2P: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14388–14402, 2022d.

Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-CLIP: Video and text adaptive CLIP via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11019–11029, 2023. doi: 10.1109/CVPR52729.2023.01062.

Yuxiang Wei, Yanteng Zhang, Xi Xiao, Tianyang Wang, Xiao Wang, and Vince D Calhoun. MoRE-Brain: Routed mixture of experts for interpretable and generalizable cross-subject fMRI visual decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.15946.

Zhikai Wei, Wenhui Dong, Peilin Zhou, Yuliang Gu, Zhou Zhao, and Yongchao Xu. Prompting segment anything model with domain-adaptive prototype for generalizable medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 533–543. Springer, 2024.

Zhiqiang Weng et al. Understanding and mitigating bias in vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://aclanthology.org/2024.emnlp-main.878.pdf.

Huangyuan Wu, Bin Li, Lianfang Tian, Chao Dong, Wenzhi Liao, and Yu Liu. CoRev2: A unified contrastive and restorative self-supervised learning method with masked prompting tuning for surface defect inspection. *IEEE Transactions on Instrumentation and Measurement*, 2025a.

Junyang Wu, Xianhang Li, Chen Wei, Huiyu Wang, Alan L. Yuille, Yuyin Zhou, and Cihang Xie. Unleashing the power of visual prompting at the pixel level. 2024a.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. *arXiv preprint arXiv:2407.21534*, 2024b.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 164–182, 2025b.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024c.

Zhengxuan Wu, Fangyu Liu, Christopher Potts, and Christopher D. Manning. A survey on visual prompting in multimodal large language models. *arXiv preprint arXiv:2409.15310*, 2024d. URL https://arxiv.org/abs/2409.15310.

Xi Xiao, Wentao Wang, Jiacheng Xie, Lijing Zhu, Gaofei Chen, Zhenqi Li, Tianyang Wang, and Min Xu. HGTDP-DTA: Hybrid graph-transformer with dynamic prompt for drug-target binding affinity prediction. In *International Conference on Neural Information Processing*, pp. 340–354. Springer Nature Singapore, 2024.

Xi Xiao, Zhengji Li, Wentao Wang, Jiacheng Xie, Houjie Lin, Swalpa Kumar Roy, Tianyang Wang, and Min Xu. Td-rd: A top-down benchmark with real-time framework for road damage detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a.

Xi Xiao, Athanasios Tsaris, Ayesha Tabassum, Jens Lagergren, Linda M. York, Tianyang Wang, and Xiao Wang. Focus: Fused observation of channels for unveiling spectra, 2025b.

Xi Xiao, Yunbei Zhang, Xingjian Li, Tianyang Wang, Xiao Wang, Yuxiang Wei, Jihun Hamm, and Min Xu. Visual instance-aware prompt tuning. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pp. 727–736, 2025c. doi: 10.1145/3672468.3672535.

Xi Xiao, Yunbei Zhang, Yanshuh Li, Xingjian Li, Tianyang Wang, Jihun Hamm, Xiao Wang, and Min Xu. Visual variational autoencoder prompt tuning, 2025d.

Xi Xiao, Yunbei Zhang, Thanh-Huy Nguyen, Ba-Thinh Lam, Janet Wang, Lin Zhao, Jihun Hamm, Tianyang Wang, Xingjian Li, Xiao Wang, Hao Xu, Tianming Liu, and Min Xu. Describe anything in medical images. In *ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*, 2025e. arXiv:2505.03804.

Xi Xiao, Yunbei Zhang, Janet Wang, Lin Zhao, Yuxiang Wei, Hengjia Li, Yanshu Li, Xiao Wang, Swalpa Kumar Roy, Hao Xu, and Tianyang Wang. RoadBench: A vision-language foundation model and benchmark for road damage understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026. arXiv:2507.17353.

Zehao Xiao and Cees GM Snoek. Beyond model adaptation at test time: A survey. *arXiv preprint arXiv:2411.03687*, 2024. URL https://arxiv.org/abs/2411.03687.

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees GM Snoek. DynaPrompt: Dynamic test-time prompt tuning. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025f.

Jiahao Xie, Alessio Tonioni, Nathalie Rauschmayr, Federico Tombari, and Bernt Schiele. Test-time visual in-context tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Qing Xie, Siyu Lin, Jingyi Xu, Jiajun Su, Zheng Chen, and Qifeng Chen. BEV-CLIP: Enhancing cross-modal alignment for BEV segmentation and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. URL https://arxiv.org/abs/2403.14492.

Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. SimDA: Simple diffusion adapter for efficient video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7827–7839, 2024. doi: 10.1109/CVPR52729.2024.00665.

Cheng Xu et al. Spectral prompt tuning for open-vocabulary segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024a.

Yifan Xu, Chaofei Liu, et al. Negative prompt driven complementary parallel structure tuning for open-world 3d object retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024b. URL https://www.ijcai.org/proceedings/2024/0166.pdf.

Yuqi Yamao, Duli Ai, Chen Yan, Xuanjing Li, Maosong Sun, and Yi Ren. IQViC: In-context, question adaptive visual compressor for VideoQA. *arXiv preprint arXiv:2408.09352*, 2024. URL https://arxiv.org/abs/2408.09352.

Siyuan Yan, Chi Liu, Zhen Yu, Lie Ju, Dwarikanath Mahapatra, Victoria Mar, Monika Janda, Peter Soyer, and Zongyuan Ge. EPVT: Environment-aware prompt vision transformer for domain generalization in skin lesion recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 249–259. Springer, 2023.

Zhibo Yan, Chuang Li, Bing Yu, et al. DTS-TPT: Dual temporal-sync test-time prompt tuning for zero-shot video action recognition. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Zhibo Yan et al. TEST-V: Test-time support-set tuning for zero-shot video classification. *arXiv preprint arXiv:2502.00426*, 2025. URL https://arxiv.org/abs/2502.00426.

Huayi Yang, Yunkang Cao, Fei Chen, Yueming Li, Zhexing Wang, and Cheng Wang. Promptable anomaly segmentation with SAM through self-perception tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. In *arXiv preprint arXiv:2310.11441*, 2023a.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. In *Advances in Neural Information Processing Systems*, 2023b.

Senqiao Yang, Jiarui Wu, Jiaming Liu, Xiaoqi Li, Qizhe Zhang, Mingjie Pan, Yulu Gan, Zehui Chen, and Shanghang Zhang. Exploring sparse visual prompt for domain adaptive dense prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5510–5518, 2024a. doi: 10.1609/aaai.v38i5.28358.

Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, Hideki Koike, et al. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. In *Advances in Neural Information Processing Systems*, 2024b.

Zefan Yang, Chenming Wu, Peng Wang, et al. BEVInstructor: Perspective-BEV prompt tuning for controllable navigation instruction generation. In *European Conference on Computer Vision (ECCV)*, 2024c.

Zesheng Ye, Chengyi Cai, Ruijiang Dong, Jianzhong Qi, Lei Feng, Pin-Yu Chen, and Feng Liu. Neural network reprogrammability: A unified theme on model reprogramming, prompt tuning, and prompt instruction. *arXiv preprint arXiv:2506.04650*, 2025.

Siqi Yin, Shaolei Liu, and Manning Wang. DDFP: Data-dependent frequency prompt for source free domain adaptation of medical image segmentation. *arXiv preprint arXiv:2505.09927*, 2025. URL https://arxiv.org/abs/2505.09927.

Seungryong Yoo, Eunji Kim, Dahuin Jung, Jungbeom Lee, and Sungroh Yoon. Improving visual prompt tuning for self-supervised vision transformers. In *International Conference on Machine Learning (ICML)*, pp. 40075–40092. PMLR, 2023.

Lang Yu, Qin Chen, Jiaju Lin, and Liang He. Black-box prompt tuning for vision-language model as a service. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2126–2134, 2023. doi: 10.24963/ijcai.2023/237.

Yuan Yuan, Yang Zhan, and Zhitong Xiong. Parameter-efficient transfer learning for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. doi: 10.1109/TGRS.2023.3236894.

Fanhu Zeng, Zhen Cheng, Fei Zhu, Hongxin Wei, and Xu-Yao Zhang. Local-prompt: Extensible local prompts for few-shot out-of-distribution detection. *arXiv preprint arXiv:2409.04796*, 2024a. URL https://arxiv.org/abs/2409.04796.

Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huang, Ying Nian Wu, and Dongfang Liu. Visual fourier prompt tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Runjia Zeng, Guangyan Sun, Qifan Wang, Tong Geng, Sohail Dianat, Xiaotian Han, Raghuveer Rao, Xueling Zhang, Cheng Han, Lifu Huang, et al. MEPT: Mixture of expert prompt tuning as a manifold mapper. *arXiv preprint arXiv:2505.18432*, 2025. URL https://arxiv.org/abs/2505.18432.

Yaohua Zha, Jinpeng Wang, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Instance-aware dynamic prompt tuning for pre-trained point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14161–14170, 2023. doi: 10.1109/ICCV51070.2023.01293.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. URL https://arxiv.org/abs/1910.04867.

Mengmeng Zhan, Zongqian Wu, Rongyao Hu, Ping Hu, Heng Tao Shen, and Xiaofeng Zhu. Towards dynamic-prompting collaboration for source-free domain adaptation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 182–188, 2024.

Chuyan Zhang, Hao Zheng, Xin You, Yefeng Zheng, and Yun Gu. PASS: Test-time prompting to adapt styles and semantic shapes in medical image segmentation. *arXiv preprint arXiv:2410.01573*, 2024a. URL https://arxiv.org/abs/2410.01573.

Hang Zhang et al. Joint vision-language social bias removal for CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1020–1031, 2023.

Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. URL https://arxiv.org/abs/2304.13785.

Qiang Zhang, Decheng Wang, and Xiao Yu. RLita: A region-level image-text alignment method for remote sensing foundation model. *Remote Sensing*, 17(10):1661, 2025b. doi: 10.3390/rs17101661.

Qingyan Zhang, Yulun Zhang, Zhang Kai, and Radu Timofte. UCIP: Universal compressed image super-resolution with dynamic prompt learning. *arXiv preprint arXiv:2407.17569*, 2024b. URL https://arxiv.org/abs/2407.17569.

Renrui Zhang, Peng Gao, Rongyao Zhang, Shijie Yang, Kunchang Li, Yixuan Lu, Han Zhang, Yu Qiao, Hongsheng Li, and Jifeng Dai. Tip-Adapter: Training-free CLIP-Adapter for better vision-language modeling. In *European Conference on Computer Vision (ECCV)*, pp. 493–510. Springer, 2022.

Shuo Zhang, Cheng Liu, Yifan Wang, et al. Positional prompt tuning for efficient 3d representation learning. *arXiv preprint arXiv:2408.11567*, 2024c. URL https://arxiv.org/abs/2408.11567.

X. Zhang et al. Prompting segment anything model with domain-adaptive prompts. In *Computer Vision – ECCV 2024 Workshops*. Springer, 2024d.

Xing Zhang, Kai Liang, Jian Wang, et al. Improving anomaly segmentation with image-aware dynamic prompts. In *Image and Graphics*. Springer, 2025c. doi: 10.1007/978-981-96-8186-0_6.

Yizhun Zhang et al. Diff-Restorer: Unleashing visual prompts for diffusion-based image restoration. *arXiv preprint arXiv:2407.03636*, 2024e. URL https://arxiv.org/abs/2407.03636.

Yunbei Zhang, Akshay Mehra, and Jihun Hamm. OT-VP: Optimal transport-guided visual prompting for test-time adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1122–1132, 2025d.

Yunbei Zhang, Akshay Mehra, Shuaicheng Niu, and Jihun Hamm. DPCore: Dynamic prompt coreset for continual test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2025e. URL https://openreview.net/forum?id=A6zDim0rQf.

Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. FedPrompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095843.

Lin Zhao, Tianchen Zhao, Zinan Lin, Xuefei Ning, Guohao Dai, Huazhong Yang, and Yu Wang. FlashEval: Towards fast and accurate evaluation of text-to-image diffusion generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8866–8875, 2024a. doi: 10.1109/CVPR52729.2024.00755.

Lin Zhao, Yushu Wu, Xinru Jiang, Jianyang Gu, Yanzhi Wang, Xiaolin Xu, Pu Zhao, and Xue Lin. Taming diffusion for dataset distillation with high representativeness. In *International Conference on Machine Learning (ICML)*, 2025a.

Shiji Zhao, Qihui Zhu, Shukun Xiong, Shouwei Ruan, Yize Fan, Ranjie Duan, Qing Guo, and Xingxing Wei. Enhancing adversarial robustness of vision language models via adversarial mixture prompt tuning. *arXiv preprint arXiv:2505.17509*, 2025b. URL https://arxiv.org/abs/2505.17509.

Zhenyu Zhao, Bing Liu, Ming Li, and Yisen Zhang. Bayesian-guided label mapping for visual reprogramming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Liang Zheng, Jie Chen, and Yunchao Wang. InsVP: Efficient instance visual prompting from image itself. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pp. 10178–10187, 2024. doi: 10.1145/3672468.3672851.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, 2022a. doi: 10.1109/CVPR52688.2022.01633.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b. doi: 10.1007/s11263-022-01653-z.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024a. URL https://openreview.net/forum?id=66Cl1H1fQr.

Shihao Zhou, Jinshan Pan, Jinglei Shi, Duosheng Chen, Lishen Qu, and Jufeng Yang. Seeing the unseen: A frequency prompt guided transformer for image restoration. In *European Conference on Computer Vision (ECCV)*, pp. 246–264. Springer, 2024b.

Yiwei Zhou, Xiaobo Xia, Zhiwei Lin, Bo Han, and Tongliang Liu. Few-shot adversarial prompt learning on vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:3122–3156, 2024c.

Junjie Zhu, Yiying Li, Ke Yang, Naiyang Guan, Zunlin Fan, Chunping Qiu, and Xiaodong Yi. MVP: Meta visual prompt tuning for few-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024a. doi: 10.1109/TGRS.2023.3348616.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Wenyu Wang, Alon Liu, and Xin Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024b.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2639–2650, 2023. doi: 10.1109/ICCV51070.2023.00234.

Xingkui Zhu, Dingkang Liang, Xingyu Jiang, Yiran Guan, Yuliang Liu, Yingying Zhu, and Xiang Bai. LayerLink: Bridging remote sensing object detection and large vision models with efficient fine-tuning. *Pattern Recognition*, 165:111583, 2025. doi: 10.1016/j.patcog.2024.111583.