Yiming LEI, Jingqi LI, Zilong LI, Yuan CAO, Hongming SHAN, 2024. Prompt learning in computer vision: a survey. *Frontiers of Information Technology & Electronic Engineering*, 25(1):42-63. https://doi.org/10.1631/FITEE.2300389

Prompt learning in computer vision: a survey

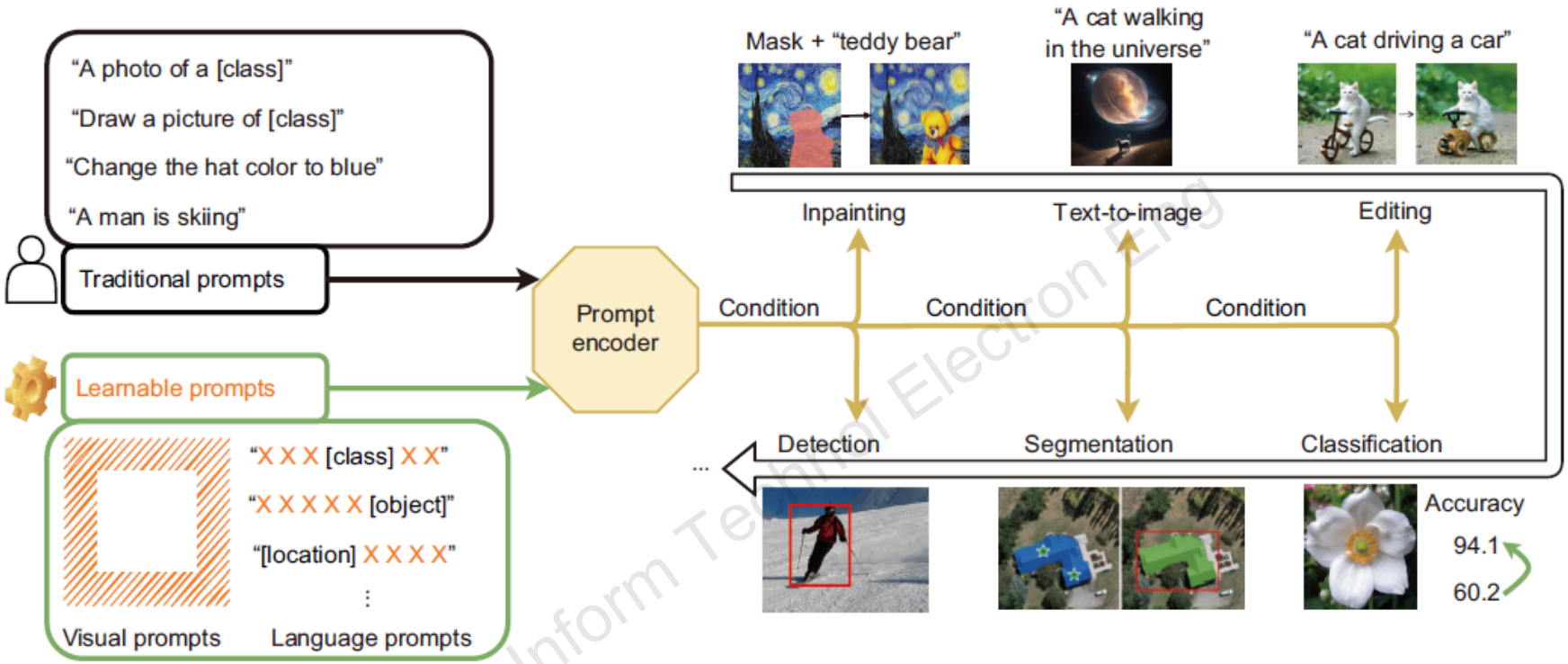
Key words: Prompt learning; Visual prompt tuning (VPT); Image generation; Image classification; Artificial intelligence generated content (AIGC)

Corresponding authors: Yiming LEI, Hongming SHAN
E-mail: ymlei@fudan.edu.cn; hmshan@fudan.edu.cn
ORCID: https://orcid.org/0000-0002-1349-7074; https://orcid.org/0000-0002-0604-3197

Motivation

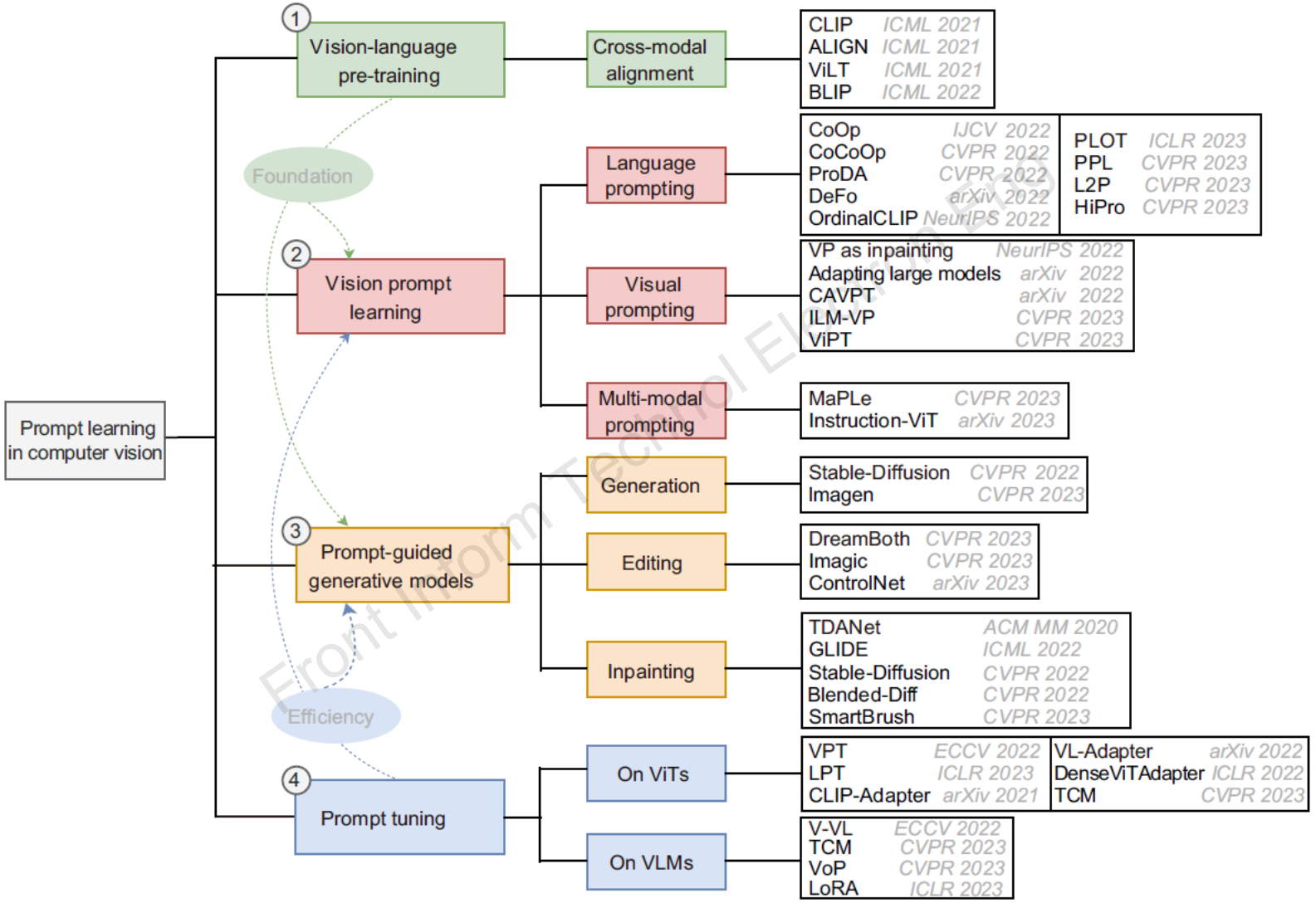
- Recently, prompt learning has attracted broad attention since the large pre-trained vision-language models (VLM) exploded such as contrastive language-image pre-training (CLIP). Prompt learning has inspired large amounts of studies and applications such as zero/few-shot learning, image restoration, and image segmentation.
- Empowered by prompt learning in both vision and language areas, artificial intelligence generated content (AIGC) becomes a hot topic for its amazing ability in generating various kinds of contents such as images, videos, and texts.
- We intend to make a comprehensive survey about prompt learning applied in computer vision (CV) community including introduction of VLM, popular prompt learning methods and prompt-guided generative models, prompt tuning, and promising directions.

Pipeline of prompt learning in CV



- Vision and language prompts act as conditions for various tasks
  - Traditional prompts are with pre-defined forms like those used in CLIP.
  - Learnable prompts are formulated as learnable parameters that will be updated during training or fine-tuning.

Overview of this survey



- The four parts interact with each other in real applications.

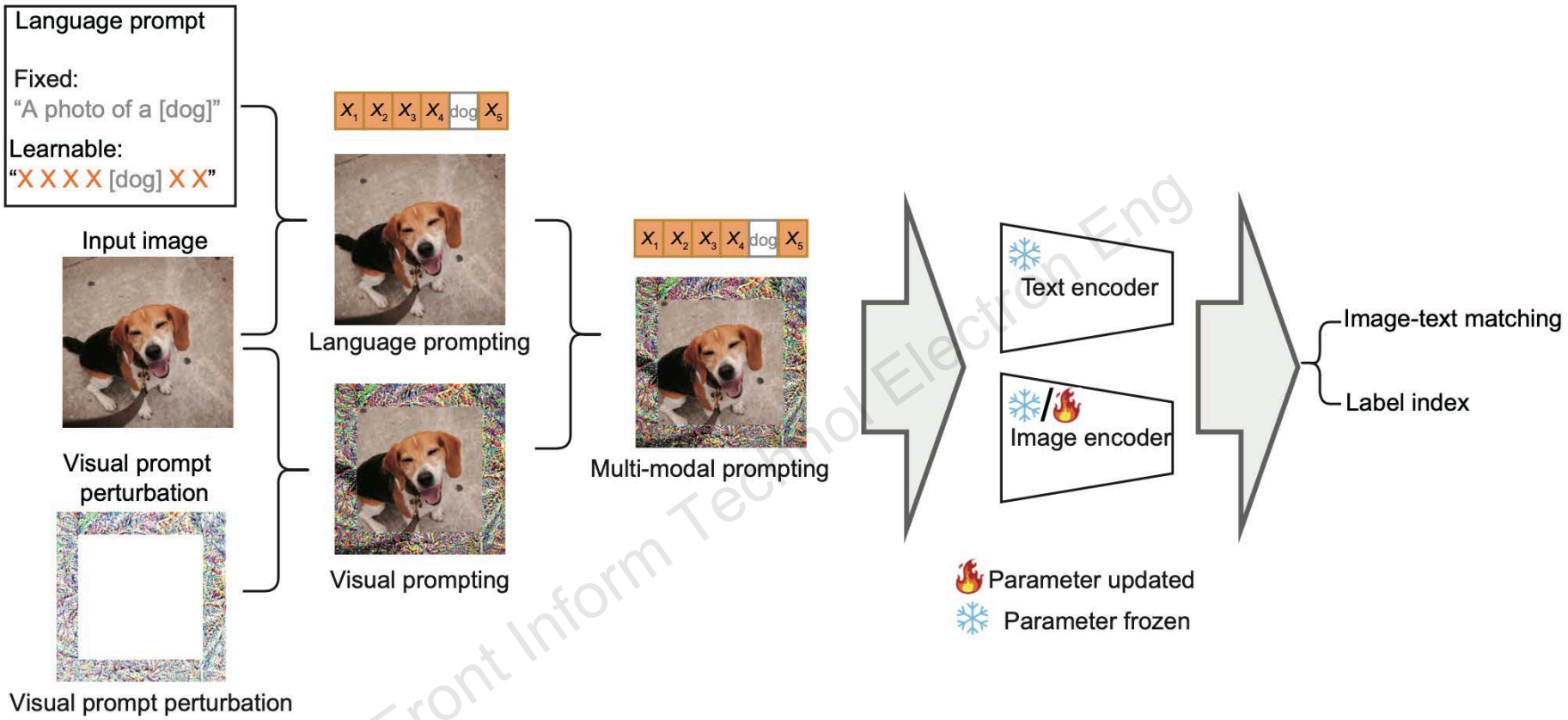
Summary of prompting methods

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

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

- Selected prompt learning methods including pre-training, downstream task adaptation, classification, and segmentation.

Hierarchical relationships prompting methods

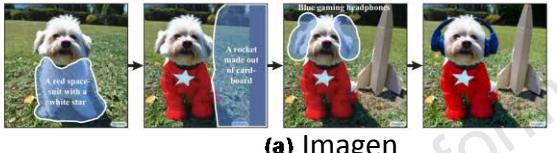


- Learnable prompts include both language and pixel tokens, and they can be combined after feeding into text and image encoders.

Prompt-guided generative models

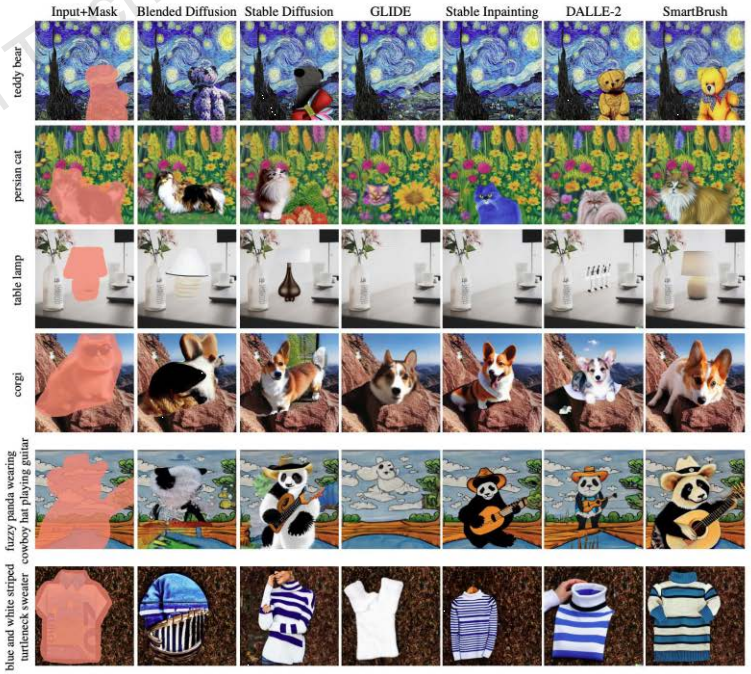
Method	Task	Language	Visual	Remark
Stable-Diffusion (Rombach et al., 2022)	Generation	✓	✓	Diffusion-based, many kinds of prompts
Imagen (Wang S et al., 2023)	Generation	✓	✗	Diffusion-based
GLIDE (Nichol et al., 2022)	Generation	✓	✗	Text-conditional image synthesis
GigaGAN (Kang et al., 2023)	Generation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

- Summary of highly-cited generative models involving diffusion models and prompt guidance.



(a) Imagen

(b) Blended-Diff



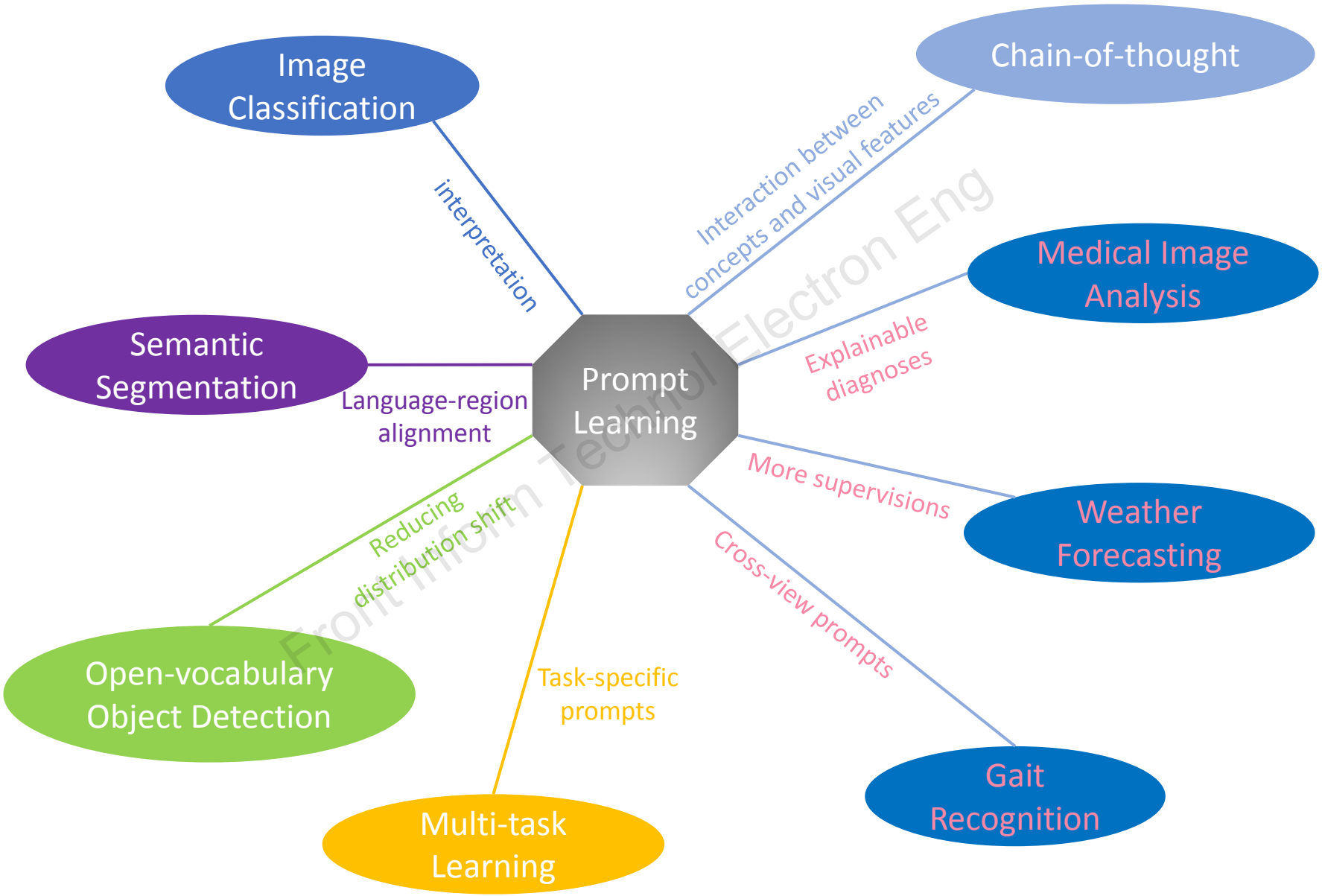
(c) SmartBrush

Prompt tuning

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

- We summarized the prompt tuning techniques into two categories: for vision Transformers (ViTs) and for VLMs.
  - For ViTs, it is imperative to study parameter-efficient tuning methods for pre-trained models.
  - For VLMs, prompt tuning can be investigated to efficiently fine-tune VLMs for transferring powerful vision and language representations.

Future directions



Conclusions

- Prompt learning strategies in image-text cross-modal frameworks have overcome shortcomings like weak generalizability of traditional prompt engineering.
- For efficiently applying large VLMs for downstream prompt-guided learning, we discussed prompt tuning methods tailored for better adaptation of large-scale ViTs.
- Prompt-guided generative models trigger a wide variety of applications such as image generation, image editing, and image inpainting.
- Prompt learning has great potential in enhancing existing studies and leading to new multi-modal directions.

Yiming LEI received the BS degree from Jinzhong University, China, in 2013, the MS degree from Qingdao University, China, in 2017, and the PhD from Fudan University. He is currently a postdoctoral researcher at the School of Computer Science, Fudan University, China, from 2022. His research interests include machine/deep learning, computer vision, and biomedical image analysis.

Jingqi LI received the BS degree from the Xidian University, Xi'an, China, in 2020. She is currently pursuing the PhD degree with the School of Computer Science, Fudan University, Shanghai, China. Her research interests include machine/deep learning, computer vision, gait recognition, and lifelong learning.

Zilong LI received the BS degree in biomedical engineering from Northeastern University, China, in 2020. He is now a PhD candidate, studying computer science at Fudan University. His primary research areas encompass machine and deep learning, medical image analysis, low-level vision, and CT reconstruction.

Yuan CAO obtained his Bachelor’s degree from Xiamen University, China, in 2013. He obtained his Master’s degree in computer science from Temple University, USA, in 2014. Subsequently, he obtained his PhD in computer software and theory from Fudan University, China, under the esteemed guidance of Professor Junping Zhang. His research interests center around deep learning and precipitation nowcasting.

Hongming SHAN received the PhD degree in machine learning from Fudan University, Shanghai, China, in 2017. From 2017 to 2020, he was a postdoctoral research associate and research scientist at Rensselaer Polytechnic Institute, Troy, NY, USA. He is currently an associate professor with the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China. His research focuses on developing machine learning algorithms for biomedical imaging.