Adversarial Attacks and Defenses in Vision-Language Pre-training: Techniques, Challenges and Opportunities

Abstract

Vision-language pretraining (VLP) has emerged as a powerful paradigm for multimodal learning. However, despite their superior capabilities, VLPs remain vulnerable to adversarial attacks by manipulating their inputs. Such attacks by undermining user trust can significantly compromise their integrity, introduce critical security vulnerabilities and highlight the importance of securing VLPs to ensure safety in various real-world multimodal applications. In the adversarial landscape of VLPs, this review aims to delve into the methodologies and implications of both adversarial attacks and defense strategies, organized by architectural considerations. Our review delves into the complexities of categorizing adversarial attack strategies, underscoring the critical need for robust defensive measures. To improve the reliability of these models, we discuss novel defense mechanisms that counter vulnerabilities. In addition, we analyze how adversarial vulnerabilities impact downstream applications. Overall, this review aims to provide a comprehensive overview of adversarial threats in VLPs and present future research directions.

1 Introduction

1.1 VLP Adversarial Attack Scenario

Vision-language pretraining models (VLPs), inspired by the success of transformers (Ngiam et al., 2011; Long et al., 2022), have revolutionized the landscape of multimodal machine learning tasks. These VLPs (Du et al., 2022), which integrate Computer Vision (CV) and Natural Language Processing (NLP), are typically pre-trained on large-scale datasets consisting of both visual and textual modalities. This paradigm enables strong multimodal understanding through rich cross-modal representations, thereby supporting a wide range of applications such as visual question answering (VQA), cross-modal retrieval, dialogue systems, image captioning, and content explanation. VLPs are increasingly deployed in security-sensitive applications, including healthcare (Finlayson et al., 2019), autonomous driving (Cao et al., 2019), and interdependent AI systems (Zhao et al., 2024), marking a key shift in the current AI technological revolution. However, VLPs remain vulnerable to adversarial attacks due to the dependence on large-scale Internet datasets for pre-training. These risks extend beyond academic interest into real-world scenarios, as VLPs are increasingly deployed in practical applications. In high-stakes scenarios, adversarial failures can lead to serious safety risks, misclassifications, or biased decision-making, thereby undermining user trust and limiting the broader adoption of VLP systems.

Adversarial attacks on VLPs (Nakano et al., 2024) have become increasingly sophisticated and diverse, targeting single or multiple modalities to induce multimodal misinterpretations and lead to erroneous predictions in complex AI systems. These findings have motivated extensive research into exploring the reliability and safety of VLPs under adversarial settings, prompting the development of attack and defense technologies. Compared to unimodal models, VLPs face unique security challenges due to their multimodal nature, as traditional adversarial attacks in CV focus on perturbing images at the pixel level and in NLP concentrate on token-level or embedding-level perturbations. Given these complexities, research into the adversarial robustness of VLPs has aimed to systematically examine the multimodal vulnerabilities arising from the joint modeling of vision and language. Understanding adversarial robustness in VLPs is essential for uncovering previously unexplored vulnerabilities and addressing emerging interdisciplinary threats. Ultimately, developing robust VLP systems requires a comprehensive examination of such security properties to enhance their reliability and safety in security-sensitive applications.

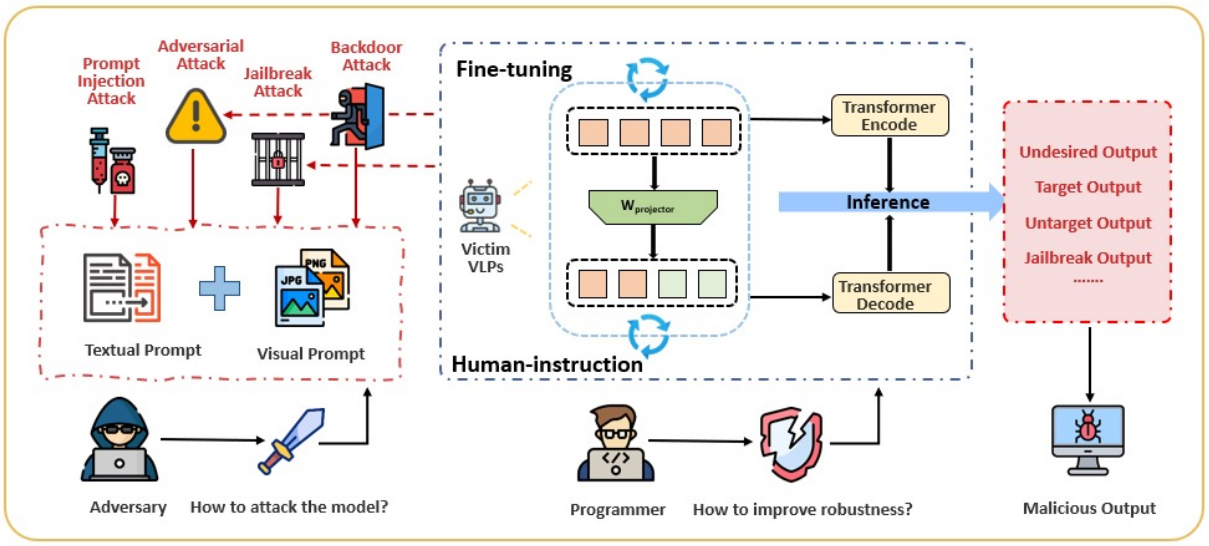

Figure 1: Overview of adversarial attack approaches against VLPs. The figure illustrates how adversaries can manipulate textual and visual inputs through prompt injection, adversarial perturbations, jailbreak attacks, and backdoor attacks.

Adversarial attacks for VLPs deliberately target inputs across different modalities to mislead models, posing a significant threat to VLP robustness. Given this threat, adversarial attacks and defenses for VLPs have received increasing research attention (Nakano et al., 2024). However, despite this growing interest, there is a lack of a systematic review that comprehensively examines these attack and defense strategies for VLPs. To the best of our knowledge, this review provides the first comprehensive analysis of adversarial exploitation in VLPs. In this review, we categorize these threats and defenses as shown in Figure 1. The attacks include prompt injection, adversarial perturbations, jailbreak attacks, and backdoor attacks. Correspondingly, defense strategies include adversarial contrastive fine-tuning, adversarial prompt tuning, backdoor defense, jailbreak defense, and safety alignment.

1.2 Contributions and Motivation for This Review

VLPs operate at the convergence of visual and linguistic processing, offering a promising paradigm for multimodal understanding and generation. They increasingly encounter adversarial threats that compromise their reliability and security across various multimodal applications. Such vulnerabilities motivate us to undertake a comprehensive investigation into adversarial robustness in VLPs. To bridge this gap, we propose a novel taxonomy that systematically categorizes adversarial attack and defense techniques based on VLP architectural perspectives. Through this work, we strive to raise awareness of AI security within the research community, while providing a stepping-stone for advancing trustworthy vision-language learning. Our main contributions are as follows:

- We propose a structured taxonomy that categorizes current VLPs based on their architectures and training objectives.
- We comprehensively summarize recent adversarial attack and defense techniques for VLPs, analyzing their effectiveness, advantages, and limitations across different model types.
- We provide an in-depth discussion of key challenges and open problems in VLP adversarial robustness, highlighting promising future research directions.

Survey	Focus Area	Core Topic	Comprehensive Attack Survey	Comprehensive Defense Survey	VLP/LVLM Specific
Shayegani et al. (Shayegani et al., 2023a)	LLM/LVLM	Attack&Defense	✓	Partial	Partial
Vatsa et al. (Vatsa et al., 2024)	VLPs	Trustworthiness	Partial	Partial	✓
Jin et al. (Jin et al., 2024)	LLM/LVLM	Jailbreaking (A&D)	Partial	✓	Partial
Liu et al. (Liu et al., 2024b)	LVLM	Attacks Only	✓	✗	✓
Dang et al. (Dang et al., 2024)	LVLMs	Interpretability	✗	✗	Partial
Zhang et al. (Zhang et al., 2024d)	LVLMs	Trustworthiness	Partial	Partial	✓
Ours	**VLP/LVLM**	**A&D**	✓	✓	✓

Table 1: Comparison between our survey and other related trustworthiness/security-related surveys in the VLPs domain.

1.3 Comparison to Existing Surveys on Adversarial Robustness in VLPs

While prior works have surveyed adversarial attacks and defenses for Large Language Models (LLMs) and Large Vision-Language Models (LVLMs), Limited work has provided a comprehensive analysis specifically focused on VLP types and their adversarial robustness. Table 1 compares our review with existing literature regarding scope, focus and coverage. We summarize the key differences as follows:

- **Shayegani et al. (Shayegani et al., 2023a):** Summarize adversarial attacks and defenses on Large language Model (LLM) and Large Vision-language Model (LVLM). They provide a taxonomy of existing attacks based on single modality, multiple modalities and additional attack types with threat model scenarios.
- **Vatsa et al. (Vatsa et al., 2024):** Review the trustworthiness of VLPs from three aspects: bias, robustness, and interpretability, with a particular focus on VLP architectures and downstream tasks.
- **Jin et al. (Jin et al., 2024):** Provide a detailed explaination of jailbreaking attacks for LLMs, categorizing seven jailbreak types and corresponding defense strategies, with brief coverage of LVLMs.
- **Liu et al. (Liu et al., 2024b):** Discuss security vulnerabilities in LVLMs, covering adversarial attacks, jailbreak attacks, prompt injection attacks, and backdoor attacks.
- **Dang et al. (Dang et al., 2024):** Focus on interpretability and explainability of LVLMs for trustworthiness, analyzing existing research from data, model, and training/inference perspectives.
- **Zhang et al. (Zhang et al., 2024d):** Present a unified benchmark for evaluating LVLM trustworthiness across five aspects: truthfulness, safety, robustness, fairness and privacy, covering various attack scenarios and bias issues.
- **Liu et al. (Liu et al., 2023b):** Provide a survey of trustworthy machine learning that reviews robustness, security, interpretability and fairness under a data-centric perspective, and extends this perspective to causal inference methods and frameworks for large pretrained models.

Based on the aforementioned works, this review distinguishes itself from existing studies in several key aspects. First, we provide a comprehensive analysis of various VLP architectures, along with their training process and targets, filling the gap in existing literature that primarily focuses on LLMs or LVLMs. Second, we systematically investigate cross-modal vulnerabilities and the unique security challenges posed by multimodal interactions, analyzing corresponding attack and defense strategies through our proposed VLP taxonomy. Third, we analyze attack and defense challenges within each category of our taxonomy. This analysis identifies key trends, highlights existing limitations, and uncovers potential solutions that have been investigated in the previous works.

We conduct a systematic literature review spanning multiple research domains, including NLP, CV, machine learning, and multimodal learning. We focus on publications from 2020 to 2025 to capture the most recent advancements in adversarial robustness for vision-language models. Our search strategy encompasses premier venues, including top-tier conferences such as NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, ACL, EMNLP, AAAI, IJCAI, and leading journals such as IEEE TPAMI, IEEE TIP, JMLR, and ACM Computing Surveys.

We further supplement this process with comprehensive searches using Google Scholar, arXiv, and major digital libraries to ensure broad coverage of relevant literature. We adopt a three-stage screening methodology: (1) title-based initial filtering to identify potentially relevant studies; (2) abstract-based assessment to evaluate direct relevance to adversarial robustness in VLP; and (3) full-text review for final inclusion decisions when abstracts provide insufficient information. This systematic approach ensures comprehensive coverage while maintaining high relevance standards. Table 2 provides definitions for all abbreviations used throughout this review.

Abbreviation	**Full Form**	**Abbreviation**	**Full Form**
3VL	Tree-Augmented Vision-Language	ACL	Adversarial Contrastive Learning
AI	Artificial Intelligence	APGD	Auto-PGD
ALIP	Adaptive Language–Image Pre-training	BaThe	Backdoor Trigger Shield
APT	Adversarial Prompt Tuning	CLAP	Contrastive Learning with Augmented Prompts
C-AVP	Class-wise Adversarial Visual Prompting	CLIP	Contrastive Language-Image Pre-Training
CNN	Convolutional Neural Network	CV	Computer Vision
ECSO	Eyes Closed and Safety On	EOS	End-of-Sequence
FAP	Few-shot Adversarial Prompt Learning	FARE	Fine-tuning for Adversarially Robust Embeddings
GAN	Generative Adversarial Networks	GPT	Generative Pre-trained Transformer
IC	Image Captioning	ITR	Image Text Retrieval
LLM	Large Language Model	LVLM	Large Vision-Language Model
MMT	Multimodal Machine Translation	MTG	Multimodal Text Generation
NLG	Natural Language Generation	NLP	Natural Language Processing
NLU	Natural Language Understanding	NLVR	Natural Language for Visual Reasoning
NoCaps	Novel Object Caption at Scale	OCR	Optical Character Recognition
ORCA	Observe–Reason–Critique–Act	OVC	Open Vocabulary Certification
PGD	Projected Gradient Descent	PSA-VLM	Progressive Safety Alignment for Vision-Language Models
RAN	Rectify Adversarial Noise	SVM	Support Vector Machine
TAPT	Test-Time Adversarial Prompt Tuning	TIJO	Trigger Inversion using Joint Optimization
UMK	Universal Master Key	VD	Visual Dialogue
VE	Visual Entailment	VG	Visual Grounding
ViT	Vision Transformer	VCR	Visual Commonsense Reasoning
VLCL	Vision-Language Contrastive Learning	VLFP	Vision-Language Fusion Pretraining
VLN	Visual Linguistic Navigation	VLP	Vision-Language Pretraining
VQA	Visual Question Answering	VR	Visual Reasoning

Table 2: Abbreviations and Definitions

As illustrated in Fig 2, the remainder of this survey is organized as follows: Section II provides a detailed review of recent advances in VLPs. Section III categorizes adversarial attacks against VLPs and delivers a thorough analysis of each category. Section IV summarizes various adversarial defensive strategies for VLPs, examining their effectiveness and limitations against different adversarial manipulation. Section V introduces vision-language downstream tasks and examines how adversarial attacks impact their performance. Finally, Section VI discusses key challenges and highlights promising future research directions for enhancing adversarial robustness in VLPs.

2 Vision-Language Pretraining Paradigms

Modern VLP frameworks have demonstrated remarkable effectiveness in integrating text and image pretraining for multimodal tasks (Mogadala et al., 2021; Long et al., 2022). This VLP paradigm excels at extracting features from complex high-dimensional data by leveraging the ability to model long-range dependencies and integrate information across modalities. VLP training strategies typically combine supervised learning with large-scale self-supervised pretraining, enhancing cross-modal representations through specialized pretraining objectives and advanced architectures. This approach enables models to learn semantic correspondences across modalities, benefiting downstream tasks while avoiding training from scratch. The current tendency in this field has been moving toward developing larger models and utilizing more extensive pretraining datasets (Chen et al., 2023b). However, existing VLPs exhibit diverse architectural frameworks and lack built-in security protections, exposing security risks under adversarial conditions. To systematically analyze

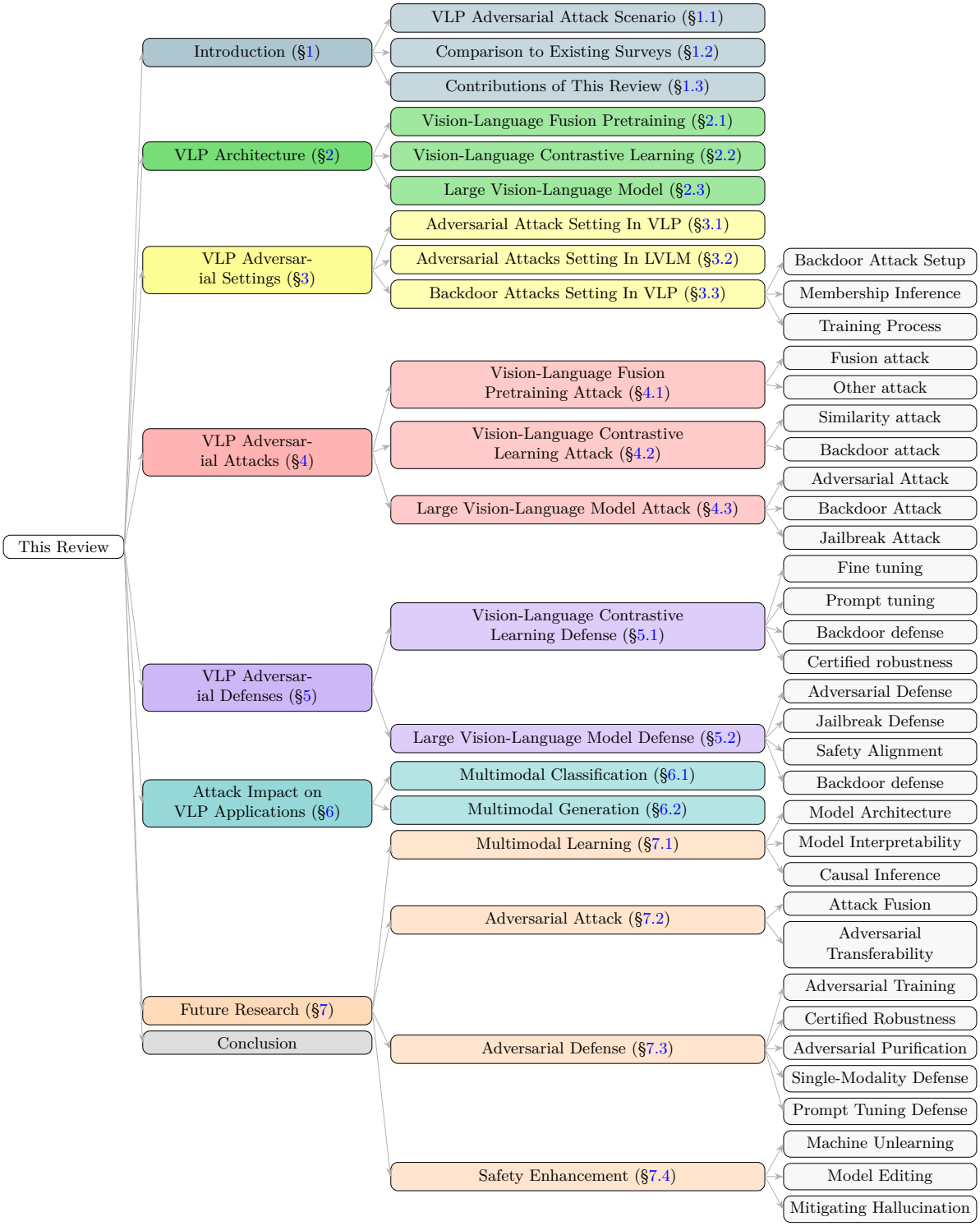

Figure 2: This review is organized into seven main sections: : (1) VLP adversarial robustness introduction, (2) VLP architectures and multimodal tactics, (3) VLP adversarial settings, (4) VLP adversarial attacks, (5) VLP adversarial defenses, (6) Downstream application impacts and (7) Future research directions.

VLP Type	Text Encoder	Vision Encoder	Fusion Scheme	Pre-training Tasks	Multimodal Datasets for Pre-training
VLFP	VisualBERT (Li et al., 2020a)	BERT	Faster R-CNN	MLM+MLM+WR+A+MFR+MRC	COCO+VG+CC+SBU
	UNITER (Chen et al., 2020)	BERT	Faster R-CNN	MLM+MRC+ITM	COCO+VG+CC+SBU
	OSCAR (Li et al., 2020b)	BERT	Faster R-CNN	MLM+MRC+ITM	COCO+VG+CC+SBU+Flickr30k
	ViLBERT (Lu et al., 2019)	BERT	Faster R-CNN	MLM+MRC+ITM+VQA	COCO+VG+VQA
	VL-BERT (Su et al., 2020)	BERT	Faster R-CNN	MLM+MRC+VQA	COCO+VG+CC+SBU
	SOHO (Huang et al., 2021)	BERT	ResNet+ViT	MLM+MV+ITM	C4+ALIGN
	Unified VLP (Li et al., 2020c)	UniLM	Faster R-CNN	MLM+MRC+VQA	COCO+VG+CC+SBU
	Pixel-BERT (Dai et al., 2020)	BERT	Faster R-CNN	MLM+seg2seg	COCO+VG+CC+SBU
	VLMo (Zhu et al., 2021)	BERT	ViT	Single stream	COCO+VG+CC+SBU
	FLAVA (Singh et al., 2021)	ViT	ViT	Single stream	LAION-400M
	ViLT (Kim et al., 2021)	BERT	ViT	Single stream	COCO+VG+CC+SBU
	BLIP (Li et al., 2022a)	BERT	ViT	Single stream	COCO+VG+CC+SBU+RedCaps
	METER (Dou et al., 2022)	BERT	ViT/ResNet	Cross-Attention Fusion	COCO+VG+CC+SBU
	BLIP-2 (Li et al., 2023a)	T5/Vicuna	ViT + Q-Former	Q-Former Projection	COCO+CC+VG+WebCorpus
VLCL	CLIP (Radford et al., 2021)	ViT	No fusion	Zero-shot Image Classification	400M web image-text pairs
	ALIGN (Jia et al., 2021)	BERT	EfficientNet	Contrastive Pretraining	1.8B web image-text pairs
	FILIP (Yao et al., 2022)	BERT	ViT	Contrastive Pretraining	CC12M
	FLIP (Li et al., 2023b)	BERT	ViT	Random Masking Pretraining	CC12M
	CLIPPO (Tschannen et al., 2023)	ViT	ViT	Pixel-based Transformer	CC12M
	MaskCLIP (Dong et al., 2023)	BERT	ViT	Self-Distillation	CC12M
	OpenCLIP (Ilharco et al., 2021)	BERT	ViT	Contrastive Learning	LAION-400M
	DeCLIP (Li et al., 2022b)	BERT	ViT	Supervised Pretraining	LAION-400M
	xCLIP (Zhou et al., 2023a)	BERT	ViT	Contrastive Learning	LAION-400M
	SLIP (Mu et al., 2022)	BERT	ViT	Contrastive Learning	LAION-400M
	Florence (Yuan et al., 2021)	BERT	ViT	Contrastive Learning	900M web image-text pairs
	BASIC (Pham et al., 2023)	BERT	ViT	Contrastive Learning	1.2B web image-text pairs
LVLM	InstructBLIP (Dai et al., 2024)	BERT	CLIP ViT	Q-Former	Custom instruction dataset
	LLaVA (Liu et al., 2023a)	LLaMA	CLIP ViT	Trainable Projection	LAION-CC-SBU + GPT-4 instructions
	MiniGPT-4 (Zhu et al., 2023)	Vicuna	ViT	Q-Former	CC3M+CC12M+SBU+LAION115M
	Flamingo (Alayrac et al., 2022)	BERT	Resampler ViT	Cross-Attention	Web image-text pairs etc.
	Openflamingo (Awadalla et al., 2023)	BERT	CLIP ViT	Cross-Attention	LAION-400M
	Multimodal-GPT (Gong et al., 2023a)	BERT	ViT	Dense Cross-Attention	LAION-400M
	PandaGPT (Su et al., 2023)	BERT	ViT	Cross-Modal Learning	LLaVa & Mini-GPT4 160k instructions
	SPHINX-X (Liu et al., 2024c)	BERT	ViT	Multimodal Learning	LAION-400M&COCO + Custom dataset
	BLIVA (Hu et al., 2024)	BERT	ViT	Multimodal Learning	COCO+LLaVA-Instruct150K etc.
	DeepSeek-VL (Liu et al., 2024a)	DeepSeek LLM	ViT	Linear Projection	DeepSeek-LLM—2T+MMC4+ShareGPT4 etc.

Table 3: Summary of VLFP, VLCL and LVLMs with architectural components, including the text encoder, vision encoder, fusion scheme, pretraining tasks and the multimodal datasets used for training.

adversarial robustness across different model types, we categorize VLP architectures from three perspectives: Vision-language Fusion Pretraining (VLFP) , Vision-language Contrastive Learning (VLCL) and LVLM, as illustrated in Fig. 3. Also this section concludes the discussion of the various VLP types listed in Table 3, highlighting that representative VLPs differ in architectural design and training objectives.

2.1 Vision-language Fusion Pretraining

VLFP represents a prevalent approach in multimodal fusion, comprising three core components: vision encoders, language encoders and multimodal fusion modules. Vision Encoder, also known as visual feature extractor, uses Convolutional Neural Network (CNN) for Grid Features (He et al., 2016), Object Detection for Region Features (Ren et al., 2015), Vision Transformer (ViT) (Dosovitskiy et al., 2021) and CLIP-ViT (Radford et al., 2021) for patch features. Language Encoder, also known as linguistic feature extractor, generates textual representations by processing input sequences with special tokens. Pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020) are typical examples of such encoders and have demonstrated significant improvements in NLP tasks. The Multimodal Fusion Module integrates the encoded image and text features within the same semantic space using the attention mechanism to enable cross-modal interactions. These architectural designs enable the complementary strengths of visual and linguistic representations, enhancing cross-modal representations to achieve remarkable success in multimodal understanding tasks.

The VLFPs have been widely adopted in various vision-language tasks, with their fusion architectures generally categorized into two common types: single-stream and dual-stream (Cho et al., 2021). The single-stream aligns with the encoding structure of the Transformer framework, to integrate visual and textual information directly concatenated into an unified representation. For instance, VisualBERT (Li et al., 2020a) proposes a flexible framework that uses object detection as visual features for image-text fusion without explicit supervision. Oscar (Li et al., 2020b) presents an object-semantics alignment approach that uses object tags recognized in images as anchor points to assemble text-tag-image triples for enhanced fusion, though this inadvertently introduces redundancy and attribute noise. VinVL (Zhang et al., 2021a) extends this idea

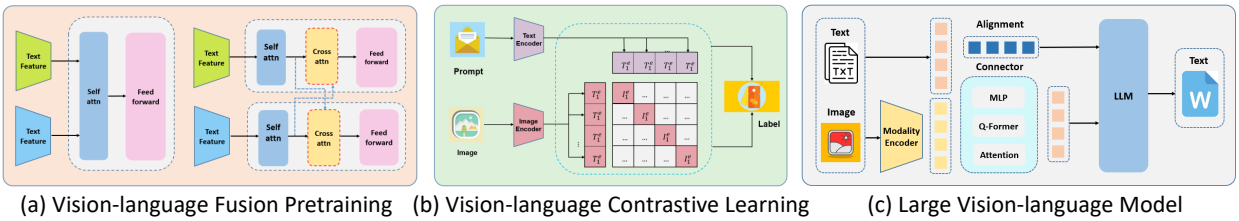

Figure 3: Overview of Vision-Language Pretraining Structures. (a) Vision-Language Fusion Pretraining – A transformer-based architecture that utilizes self-attention and cross-attention mechanisms to effectively integrate text and image features. (b) Vision-Language Contrastive Learning – A contrastive learning framework where separate text and image encoders project inputs into a shared latent space, optimizing alignment between corresponding pairs. (c) Large Vision-Language Model – A multimodal framework that incorporates modality encoders, alignment mechanisms, and an LLM to process multimodal inputs and generate coherent text outputs.

by developing a new object detection model that incorporates the object attribute into image encoding. This learning mode has enhanced visual representations at the cost of increased model size and computational overhead. Instead of replying on pre-trained object detectors, ViLT (Kim et al., 2021) simplifies the architecture by directly integrating image features through computationally efficient ViT encoder with BERT for linguistic features. UNITER (Chen et al., 2020) employs conditional masking strategies to mask either the image or text (one modality at a time). This is opposed to the joint random masking on both modalities during pretraining. The Dual-Stream processes visual and linguistic features separately using independent transformer encoder layers. These features are then integrated through a cross-attention layers for cross-modality interactions. LXMERT (Tan & Bansal, 2019) applies self-attention layers to each modality independently (e.g., nine layers for text and five layers for image) before fusing features through 5 cross-attention layers. METER (Dou et al., 2022) comprehensively explores various multimodal fusion strategies in vision-language downstream tasks, evaluating fused performance across on merged-attention and cross-attention mechanisms, as well as encoder-only and encoder-decoder architectures. The BLIP series (Li et al., 2022a; 2023a) adopts flexible transformer frameworks to support both vision-language discriminative and generative tasks. This unified framework effectively utilizes encoder-decoder structures to handle diverse tasks involving noisy web data. Both VLFP architectural paradigms by leveraging complementary information across modalities have driven significant performance improvements, including ViLBERT (Lu et al., 2019), VL-BERT (Su et al., 2020), FLAVA (Singh et al., 2022) and ALBEF (Li et al., 2021). This has been demonstrated to boost performance on classification tasks and further push the performance boundaries of VLFP.

2.2 Vision-language Contrastive Learning

VLCL leverages contrastive learning as a general-purpose framework for learning from unlabeled paired image-text data. Such as CLIP (Radford et al., 2021) has demonstrated remarkable performance in large-scale pretraining, utilizing 400 million image-text pairs collected from the Internet. CLIP shows that how natural language supervision can effectively learn visual representations without requiring manually annotated labels. This methodology involves projecting textual and visual features into a shared multimodal embedding space using contrastive objectives. The contrastive learning process applies a similarity measure that minimizes the distance between matched image-text pairs while maximizing the distance between unmatched pairs, enabling the model to learn semantic correspondences across modalities. Notably, CLIP achieves impressive zero-shot and few-shot performance in image classification and image-text retrieval tasks without multimodal fusion. The learned representations are sufficiently robust to generalize across a wide range of vision-language downstream tasks, eliminating the need for task-specific fine-tuning. Adaptive Language–Image Pre-training (ALIP) (Yang et al., 2023a) enhances CLIP against noisy web data by generating synthetic captions as complementary supervision. It employs a bi-path architecture with adaptive gating mechanisms Language Consistency Gate and Description Consistency Gate dynamically to dynamically ad-

just weight for training. The contrastive loss improves pre-training efficiency and effectively alleviates the impact of noise data.

Specifically, ALIGN (Jia et al., 2021) expands a noisy dataset to enhance contrastive vision-language representation learning. This approach follows utilizing simple frequency-based filtering without requiring on expensive filtering or post-processing steps. CyCLIP (Goel et al., 2022) introduces cross-modal and in-modal consistency regularizers to constrain similarity relationships and enforce geometric alignment between image and text embeddings. This design improves the consistency of image-text representations and better preserves semantic hierarchies across modalities for improved performance. FILIP (Yao et al., 2022) extends CLIP by introducing token-wise maximal similarity between visual patches and linguistic tokens to guide the contrastive objective. The proposed method successfully captures finer-grained vision-language representations while maintaining efficiency in large-scale training and inference. FLIP (Li et al., 2023b) further optimizes CLIP training by adopting a high-ratio random masking strategy for visible patches, achieving a favorable trade-off between accuracy and training time while reducing computational cost without performance degradation. CLIPPO (Tschannen et al., 2023) unifies image and text representations by employing a pure pixel-based ViT, eliminating a tokenizer and pre-defined text-specific embeddings. The model learns both natural images and rendered text to images using contrastive learning optimization. This optimization halves the parameters compared to dual-tower CLIP while maintaining comparable performance. MaskCLIP (Dong et al., 2023) exploits masked self-distillation to improve local image patch learning, focusing on text-related representation and enhancing the text encoder via masked language modeling. These authors highlight that distilling representations from local semantics contribute to improve global semantic transferability. Following these advances, subsequent VLCL models such as OpenCLIP (Ilharco et al., 2021), DeCLIP (Li et al., 2022b), xCLIP (Zhou et al., 2023a), SLIP (Mu et al., 2022), Florence (Yuan et al., 2021) and BASIC (Pham et al., 2023) have exhibited the continuous evolution of CLIP-style models.

2.3 Large Vision-language Model

Building on the success of auto-regressive modeling techniques in LLMs, there has been widespread attention devoted to the development of LVLM (Zhang et al., 2024a; Caffagni et al., 2024). The prevalent strategy involves integrating visual encoders with LLMs, where visual features as additional inputs with linguistic features to improve multimodal generation capabilities. The effectiveness of learning algorithms is determined by both novel training methods and the efficiency of the model architecture (Liu et al., 2024a). The fundamental structure of LVLMs typically consists of three components: a visual encoder, a modality connection module, and an LLM backbone. The visual encoder, commonly adapted from CLIP ViT (Radford et al., 2021), converts images into visual patches as supplementary inputs. The connection module aligns these visual patches with the linguistic tokens in LLM, guaranteeing that the LLM to effectively processes the visual information. Then various modality alignment approaches have been developed, including linear layers projection (Liu et al., 2024e), Multilayer Perception (MLP) (Liu et al., 2024d), cross-attention(Alayrac et al., 2022), adapters (Zhang et al., 2024c), Querying Transformer (Q-Former) (Li et al., 2023a), and Dense Connector(Yao et al., 2024). In addition, prominent examples such as GPT-4 (Achiam et al., 2023) demonstrates significant advancements in Natural Language Understanding (NLU) and Natural Language Generation (NLG) through effective modality alignment strategies. In the end, LVLMs exhibit substantial generation capabilities in response to human instructions and have advanced significantly across various complex vision-language tasks, positioning them as a key trend in multimodal AI. Yet, LVLMs are huge model size and computationally expensive, which pose challenges in inference efficiency and low resource consumption.

InstructBLIP (Dai et al., 2024) extends BLIP-2 (Li et al., 2023a) for general-purpose vision-language instruction tuning by fine-tuning the Q-Former while keeping both the image encoder and LLM frozen. This design converts visual features into a soft prompt via a linear projection to LLM. LLaVA (Liu et al., 2024e) simplifies the sophisticated architecture by employing a pre-trained CLIP ViT connected to LLaMA (Touvron et al., 2023) while connecting to a trainable projection matrix for cross-modal representations. MiniGPT-4 (Zhu et al., 2023) consists of a frozen visual encoder ViT with a Q-Former network to align the visual and linguistic information. It requires training a linear projection layer to the frozen Vicuna (Zheng et al., 2023) language model. Additionally, Flamingo (Alayrac et al., 2022) introduces the perceiver resampler to

generate fixed visual tokens and use gated cross-attention for token-level fusion, enhancing few-shot learning adaptation. OpenFlamingo (Awadalla et al., 2023) builds on this concept using frozen CLIP ViT and open-source language model decoders, while training the cross-attention fusion modules for auto-regressive language modeling. Multimodal-GPT (Gong et al., 2023a) proposes a unified instruction template for vision-language data, facilitating improved comprehension of human instructions. This model has been trained to exhibit strong dialogue capabilities for human interactions and is fine-tuned in a parameter-efficient manner based on Openflamingo. Collectively, LVLMs have contributed to the advancement of vision-language tasks through architectural innovations. Subsequent investigations such as PandaGPT (Su et al., 2023), SPHINX-X (Liu et al., 2024c) and BLIVA (Hu et al., 2024), continue to expand the frontiers of cross-modal learning with LLMs.

2.4 Summary of Vision and Language Pretraining

Despite significant achievements, current VLP research primarily focuses on improving performance on benchmark datasets, with limited attention paid to adversarial robustness considerations. Consequently, existing evaluations often lack comprehensive robustness assessments, leaving potential vulnerabilities unexamined in real-world deployment scenarios. Research on adversarial robustness in VLPs remains in its early stages, as many critical components required to enhance their reliability have yet to be developed. This challenge has sparked growing interest in exploring the resilience of VLPs for robust reasoning under adversarial conditions. We believe that this focus on VLP adversarial robustness constitutes a crucial research direction for advancing trustworthy VLP systems.

3 VLP Adversarial Settings

3.1 Adversarial Attacks Setting in VLP

Traditional adversarial attacks primarily focus on single-modality models, where the input space is limited to visual, textual or other individual modalities. Adversarial perturbations in one modality can propagate and exacerbate misalignments in multimodal systems, leading to unexpected behavior. While single-domain models present unique input-output relationship vulnerabilities, multimodal systems introduce new ones: perturbations in one modality can propagate and induce misalignments across modalities in multimodal systems, potentially leading to unexpected behaviors.

Adversarial attacks on VLPs involve manipulating the original visual and textual inputs to create deceptive multimodal examples that can mislead VLP classification or generation models. Typically, such attacks introduce carefully designed perturbations to the image, the text, or both, with the goal of inducing incorrect predictions or target outputs while preserving the overall semantic content of the original inputs. The multimodal attack optimization objective is formulated as follows:

$$\delta_{\mathbf{v}}^*, \delta_{\mathbf{t}}^* = \underset{\|\delta_{\mathbf{v}}\| \leq \epsilon_{\mathbf{v}}, \|\delta_{\mathbf{t}}\| \leq \epsilon_t}{\arg\max} \mathcal{L}(\mathbf{x_v} + \delta_{\mathbf{v}}, \mathbf{x_t} + \delta_{\mathbf{t}}, \mathbf{y}) \tag{1}$$

where $\delta_{\mathbf{v}}$ represents the perturbation added to the image input $\mathbf{x_v}$ and $\delta_{\mathbf{t}}$ is the perturbation added to the text input $\mathbf{x_t}$. The constraints $\epsilon_{\mathbf{v}}$ and $\epsilon_{\mathbf{t}}$ are the set of allowed maximize perturbations. The problem can be abstracted as searching for perturbations to fool $\mathbf{y}$ the expected output, while maximizing the loss function $\mathcal{L}(\cdot)$ when applying $(\delta_{\mathbf{v}}^*, \delta_{\mathbf{t}}^*)$ perturbations to effectively deceive the model.

Adversarial attacks on VLPs can take various forms depending on the targeted modality. When attacks focus solely on textual inputs while the image remains unchanged, the attack primarily targets the language encoder or prompt-conditioning mechanisms by employing token-level manipulations, embedding perturbations or prompt injection techniques. Conversely, when perturbations are applied exclusively to visual inputs with fixed textual inputs, these attacks disrupt the visual encoder to impact cross-modal interaction, typically utilizing imperceptible adversarial pixel or patch modifications. In more complex scenarios, VLP attacks can simultaneously manipulate both modalities to exploit vulnerabilities in cross-modal alignment (Wang et al., 2024c). Such multimodal attacks have been shown to be more effective than those targeting a single modality.

3.2 Adversarial Attacks Setting in LVLM

LLMs such as GPT-3 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023), rely on probabilistic mechanisms to generate coherent and contextually relevant sequences of text. For a given input embedding sequence $\mathbf{H}^t_{1:n}$ derived from textual input $\mathbf{x}^t$, the joint probability of generating an output sequence $\mathbf{y}$ is expressed as:

$$p(\mathbf{y} \mid \mathbf{H}_{1:n}) = \prod_{i=1}^{m} p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1}, \mathbf{H}^t_{1:n}), \quad \text{where } \mathbf{H}^t_{1:n} = \text{Encoder}(\mathbf{x}^t) \tag{2}$$

where $\mathbf{y}_i$ denotes the probability of generating the i-th token given on the previously generated tokens $\mathbf{y}_{1:i-1}$ and the input representation $\mathbf{H}^t_{1:n}$. This sequential modeling approach leverages the model to generate linguistically coherent and contextually appropriate text through autoregressive left-to-right decoding.

Adversarial attacks on LLMs primarily aim to manipulate this probability by introducing carefully crafted perturbations to the input embeddings. Let $\tilde{\mathbf{H}}_{1:n}$ represents the adversarially perturbed embeddings, then the optimization for generating adversarial embeddings is formalized as:

$$\tilde{\mathbf{H}}^{\mathbf{t}}_{1:n} = \underset{\tilde{\mathbf{H}}^{\mathbf{t}}_{1:n} \in \mathcal{A}(\mathbf{H}^{\mathbf{t}}_{1:n})}{\arg\max} \; -\log p(\mathbf{y}^\star \mid \tilde{\mathbf{H}}^{\mathbf{t}}_{1:n}) \tag{3}$$

where $\mathbf{y}^\star$ is the adversarially targeted output sequence. $\mathcal{A}(\mathbf{H}_{1:n})$ defines the allowable perturbation space constrained by a norm $\|\tilde{\mathbf{H}}^{\mathbf{t}}_{1:n} - \mathbf{H}^{\mathbf{t}}_{1:n}\| \leq \epsilon$, where ϵ denotes the maximum allowable perturbation magnitude.

When attacking aligned models, the adversary may also target latent reward functions, which are modeled to align the output with human preferences. The reward function $\mathcal{R}^*(\mathbf{y} \mid \mathbf{H}^t_{1:n})$ evaluates the alignment of the output sequence $\mathbf{y}$ with the desired criteria. In adversarial settings, the goal is to minimize the reward while inducing harmful or undesired outputs:

$$\mathbf{y}^\star = \min \mathcal{R}^*(\mathbf{y} \mid \tilde{\mathbf{H}}^t_{1:n}) \tag{4}$$

In practice, textual adversarial attacks on LLMs can manifest as substituting or rearranging words (e.g., synonyms or homophones), inserting subtle typos that manipulate token embeddings or crafting prompts designed to lead the language model to generate harmful or unintended outputs. As LLMs are increasingly applied in various scenarios such as code generation, chat-based assistant tasks and policy-making suggestions, adversarial text perturbations present serious risks by injecting misinformation, hateful content or malicious instructions.

LVLMs integrate visual-linguistic modalities to perform multimodal tasks, relying on the alignment between visual and textual embeddings to generate responses. Therefore, LVLMs introduce new vulnerabilities (jailbreak attacks) that adversaries can exploit by disrupting modality alignment and cross-modal perturbations to unexpected behaviors.

Let $\mathbf{x}^{\mathbf{v}}$ denote the input image and $\mathbf{x}^{\mathbf{t}}$ denote the input text sequence consisting of n tokens. The LVLM encodes the image using a visual encoder $g(\cdot)$, producing visual feature vectors $\mathbf{Z}_{\mathbf{v}} = g(\mathbf{x}^{\mathbf{v}}) \in \mathbb{R}^{d \times k}$, where d is the feature dimension and k is the number of image tokens. The input text is tokenized and embedded to obtain language embeddings $\mathbf{H}^{\mathbf{t}}_{1:n} \in \mathbb{R}^{d \times n}$. To fuse modalities, a learnable projection matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ is used to map the visual features into the language embedding space:

$$\mathbf{H}^{\mathbf{v}}_{1:k} = \mathbf{W} \cdot \mathbf{Z}_{\mathbf{v}} \tag{5}$$

The LVLM then conditions on the concatenated multimodal sequence $[\mathbf{H}^{\mathbf{v}}_{1:k}, \mathbf{H}^{\mathbf{t}}_{1:n}]$ to generate an output sequence $\mathbf{y}$. The joint probability of the output sequence is given by:

$$p(\mathbf{y} \mid [\mathbf{H}^{\mathbf{v}}_{1:k}, \mathbf{H}^{\mathbf{t}}_{1:n}]) = \prod_{i=1}^{m} p(\mathbf{y}_i \mid \mathbf{y}_{1:i-1}, [\mathbf{H}^{\mathbf{v}}_{1:k}, \mathbf{H}^{\mathbf{t}}_{1:n}]) \tag{6}$$

where m is the output length. Adversarial attacks on LVLMs aim to manipulate this joint probability by introducing perturbations to both the visual and textual embeddings, denoted as $\tilde{\mathbf{H}}^{\mathbf{v}}_{1:k}$ and $\tilde{\mathbf{H}}^{\mathbf{t}}_{1:n}$, respectively. These perturbations are designed to maximize an adversarial loss $\mathcal{L}^{adv}$, defined as:

$$\mathcal{L}^{adv}([\tilde{\mathbf{H}}^{\mathbf{v}}_{1:k}, \tilde{\mathbf{H}}^{\mathbf{t}}_{1:n}]) = -\log p(\mathbf{y}^{\star} \mid [\tilde{\mathbf{H}}^{\mathbf{v}}_{1:k}, \tilde{\mathbf{H}}^{\mathbf{t}}_{1:n}]), \tag{7}$$

where $\mathbf{y}^{\star}$ represents the adversarial attack goal. The adversarial embeddings are obtained by solving:

$$[\tilde{\mathbf{H}}^{\mathbf{v}}_{1:k}, \tilde{\mathbf{H}}^{\mathbf{t}}_{1:n}] = \underset{[\tilde{\mathbf{H}}^{\mathbf{v}}_{1:k}, \tilde{\mathbf{H}}^{\mathbf{t}}_{1:n}] \in \mathcal{A}([\hat{\mathbf{H}}^{\mathbf{v}}_{1:k}, \hat{\mathbf{H}}^{\mathbf{t}}_{1:n}])}{\arg\min} \mathcal{L}^{adv}([\tilde{\mathbf{H}}^{\mathbf{v}}_{1:k}, \tilde{\mathbf{H}}^{\mathbf{t}}_{1:n}]) \tag{8}$$

where $\mathcal{A}$ defines the adversarial search space.

In addition to input perturbations, recent work (Li et al., 2024d; Pantazopoulos et al., 2024) indicates that adversaries can also exploit vulnerabilities in cross-modal attention and alignment mechanisms. For instance, slightly altering textual prompts may redirect attention to irrelevant or misleading regions of the image, effectively causing confusion in visual grounding. Likewise, small image perturbations can fool the text embedding alignment and cause unintended generation. Mitigating such risks often involves adversarial training across multimodal data, gradient masking strategies, or more advanced approaches like adversarial contrastive learning that jointly penalize inconsistent embeddings across vision and language components.

3.3 Backdoor Attacks Setting In VLP

Backdoor attacks are a particularly insidious category of adversarial threats that aim to embed hidden "triggers" into a model during training. Unlike conventional adversarial perturbations (cf. Section **??**), which typically operate at test-time by adding a small norm-bounded δ to the input, backdoor attacks tamper with the *training process* itself to implant malicious functionality. The compromised model behaves normally on benign inputs but produces attacker-specified outputs when a specific trigger pattern is present.

Backdoor Attack Setup. Formally, let $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}$ denote the original (benign) training set, where x_i can be an image, text snippet, or a multimodal input (e.g., image-text pair). In a backdoor scenario, the adversary constructs a *poisoned* training set $\mathcal{D}_{\text{poison}}$ by injecting pairs $(x_i + \tau, y_{\text{tgt}})$ for a subset of training samples, where:

- τ is a *trigger pattern* added to x_i. This trigger could be a small visual patch for images or a keyword token for text.
- y_{tgt} is the *target label* chosen by the adversary.

Thus, the overall training set becomes

$$\mathcal{D}'_{\text{train}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{poison}} \tag{9}$$

When the model is trained on $\mathcal{D}'_{\text{train}}$, it learns to associate the presence of τ with the malicious label y_{tgt}, while still performing well on clean inputs x. At inference time, any test input $x_{\text{test}} + \tau$ will trigger the model to predict y_{tgt}, effectively bypassing the model's legitimate decision boundary.

Membership Inference Attacks via Backdoors. Beyond forcing targeted misclassification, backdoors can also facilitate membership inference attacks, especially in large-scale VLPs such as CLIP. Membership inference attacks (Shokri et al., 2017) attempt to ascertain whether a particular sample $\tilde{x}$ was part of the training set. In a backdoored model, the presence of a hidden trigger τ—or knowledge of how the model responds to τ—can amplify such privacy leakage. Concretely, an adversary can:

- **Probe the Model's Trigger Response:** For a candidate input $\tilde{x}$, the adversary adds the backdoor trigger τ to form $\tilde{x} + \tau$. If the model's output exhibits a strong confidence for the target label y_{tgt}, it may suggest $\tilde{x}$ was part of (or closely related to) the poisoned set used to implant the trigger.

- **Exploit Latent Representations:** CLIP model family map images and text into a shared embedding space. If a backdoor trigger modifies these embeddings in a distinguishable manner only for samples used during training (or those that share semantic similarities), the adversary can detect membership by analyzing embedding shifts or gradients.

These inferences pose serious privacy risks, as merely identifying which data points were used in the training process can compromise data confidentiality or intellectual property rights.

Training Process. From a defensive standpoint, mitigating backdoor attacks demands careful vetting of each stage in the training pipeline:

- *Data Filtering and Sanitization:* Statistical methods can be employed to detect outliers or suspicious patterns among training samples (Chen et al., 2018). In vision-language settings, cross-modal consistency checks (e.g., verifying image-text alignment) may help identify poisoned instances.
- *Robust Training Protocols:* Techniques such as gradient aggregation (e.g., multi-Krum, median-based) or regularization methods can suppress the influence of outlier updates in federated or distributed training.
- *Trigger-Response Inspection:* Post-training audits can probe the model's response to various candidate triggers. If the model consistently outputs a specific label y_{tgt} upon detecting a certain pattern, this suggests a potential backdoor.
- *Certifiable Defenses:* Recent work explores certifiable robustness against backdoors, aiming to provide theoretical guarantees that a model's predictions cannot be flipped by a single, small trigger (Weber et al., 2023).

As VLPs continue to grow in size and complexity, defending against backdoor threats becomes increasingly challenging. Huge datasets and long training pipelines offer adversaries ample opportunities to embed triggers that evade standard inspections. For example, backdoor attacks remain challenging due to their stealthy nature and minimal impact on clean accuracy. In LVLMs, where multiple modalities and massive training sets are involved, adversaries have substantial freedom to camouflage triggers. Moreover, the multimodal nature of these models can mask anomalous behaviours: a subtle visual patch might be overlooked when textual features dominate, or a single text token might be buried in large corpora of data. Consequently, systematic defensive measures that check cross-modal consistency and scrutinize both the final classifier layer and the embedding space are paramount. In turn, defenders must adopt a holistic view of adversarial robustness, unifying strategies against both classical perturbation-based attacks and stealthy backdoor injection, while ensuring membership privacy is not compromised.

3.4 Summary of Adversarial Setting on VLPs

VLPs play a vital role in learning cross-modal representations and enhancing semantic understanding. However, they are vulnerable to adversarial attacks where subtle adversarial perturbations across modalities can cause significant deviations in model behavior. As such attacks can disrupt the intricate interplay between visual and textual representations, they present significant challenges for multimodal adversarial security in the rapidly evolving landscape of VLPs. As large-scale VLPs are increasingly deployed in real-world applications, understanding their adversarial vulnerabilities is crucial to prevent system compromise and harmful outputs for ensuring safe deployment.

4 Adversarial Attacks for VLPs

VLPs exhibits unique multimodal vulnerabilities that differs single-source adversarial exploitation. Through subtle perturbations applied to either visual or textual inputs or both, adversarial attacks on VLPs are categorized into single-modal and multimodal strategies, reflecting the multifaceted nature of the threat landscape. These attacks can significantly impact critical applications including VQA, Image Captioning,

Cross-Modal Retrieval and Visual Reasoning. Understanding these attacks is crucial for exploring adversarial robustness and highlighting security-critical applications of VLPs. This section reviews adversarial attacks on targeting various VLPs, as summarized in Table 4.

Attack Method	Target	Modality	Strategy	Key Contribution
Model Type: VLFP (Vision-Language Fusion Pretraining)				
Co-Attack (Zhang et al., 2022)	VLFP	Image + Text	Gradient based + Token Replacement	Combines perturbations across modalities
Wang et al. (2024g)	VLFP	Image + Text	Gradient based + Gumbel-Softmax Sampling	Enables joint multimodal optimization
SGA (Lu et al., 2023)	VLFP	Multimodal	Set-Level Data Augmentation	Guides adversarial pair generation
VLATTACK (Yin et al., 2023)	VLFP	Image + Text	Iterative Refinement	Refines cross-modal perturbations iteratively
TMM (Wang et al., 2024a)	VLFP	Multimodal	Attention-Directed Perturbation	Exploits modality-consistent features
Gao et al. (2025)	VLFP	Multimodal	Intersection Diversification	Enhances transferability via diversity balancing
OT-Attack (Han et al., 2023)	VLFP	Image + Text	Optimal Transport	Enhances adversarial quality
Sa-Attack (He et al., 2023)	VLFP	Image + Text	Self-Augmentation	Diversifies multimodal inputs
Shirnin et al. (Shirnin et al., 2024)	VLFP	Image + Text	Multi-Technique Perturbation Analysis	Compares robustness under multimodal attacks
Model Type: VLCL (Vision-Language Contrastive Learning)				
AdvCLIP (Zhou et al., 2023b)	VLCL	Image	Adversarial Patch	Minimizes embedding similarity
ETU (Zhang et al., 2024b)	VLCL	Image + Text	Universal Perturbation	Enhances cross-modal transferability
FGA (Zheng et al., 2024a)	VLCL	Image + Text	Embedding Disruption	Disrupts cross-modal alignment
Carlini & Terzis (2022)	VLCL	Image + Text	Backdoor Patch Injection	Triggers model failure with minimal poisoning
Zheng et al. (2024b)	VLCL	Image + Text	Universal Perturbation	Builds transferable global perturbations
BadCLIP (Bai et al., 2024)	VLCL	Image + Text	Trigger-Aware Context Generator	Injects coordinated backdoors
Ko et al. (2023)	VLCL	Image + Text	Membership Inference	Leverages cosine similarity thresholds
RoCLIP (Yang et al., 2023b)	VLCL	Image + Text	Backdoor Trigger + Adversarial Caption	Triggers targeted misclassification without affecting benign inputs
Model Type: LVLM (Large Vision-Language Models)				
Zhao et al. (2023)	LVLM	Image	Transfer + Query-Based Attack	Transfers adversarial visual prompts across LVLMs
Cui et al. (2024)	LVLM	Image + Text	Targeted Perturbation	Balances vision-text losses under multiple settings
Schlarmann & Hein (2023)	LVLM	Image	APGD	Contextual prompts for resilience
VT-Attack (Wang et al., 2024i)	LVLM	Image	Feature Disruption	Disrupts global visual semantics
SparseMA (Yu et al., 2023)	LVLM	Image + Text	Sparse Perturbation	Identifies vulnerabilities in discrete space
Verbose Images (Gao et al., 2024)	LVLM	Image + Text	Temporal Weight Adjustment	Maximizes sequence length with uncertainty control
Zong et al. (2024b)	LVLM	Multimodal	Permutation Sensitivity	Reveals vulnerabilities in answer sets
Li & Zhang (2023)	LVLM	Image + Text	Visual-Textual Backdoor Trigger	Triggers malicious output via multimodal inputs
Chen & Wang (2023)	LVLM	Image + Text	Stealthy Trigger Insertion	Injects hidden backdoors across modalities
BadVLMDriver (Ni et al., 2024)	LVLM	Image + Text	Physical Backdoor Attack	Injects backdoors in autonomous driving
Shadowcast (Xu et al., 2024a)	LVLM	Image + Text	Stealthy Data Poisoning	Inserts poisoned multimodal pairs via paraphrasing
AnyDoor (Lu et al., 2024)	LVLM	Image + Text	Test-Time Backdoor Injection	Injects backdoors via universal test images
Wang et al. (2024e)	LVLM	Text + Image	Jailbreak Survey	Overviews jailbreak methods, benchmarks, and defenses
Carlini et al. (2023)	LVLM	Image + Text	Adversarial Prefixes	Uses visual-text prompts to trigger harmful output
Qi et al. (2024)	LVLM	Image	Visual Adversarial Jailbreak	Evades safety alignment with visual prompts
Wang et al. (2024c)	LVLM	Image + Text	Dual Optimization (UMK)	Jointly attacks prefix/suffix with high success rate
Jailbreak in Pieces (Shayegani et al., 2023b)	LVLM	Image + Text	Adversarial Image Injection	Generates unsafe responses via images
Infectious Jailbreak (Gu et al., 2024)	LVLM	Multimodal	Multi-Agent Interaction	Cascades jailbreaks across agents
Pantazopoulos et al. (2024)	LVLM	Image + Text	Visual Prompt + Evaluation Framework	Shows higher jailbreak risk in visually tuned LVLMs
HADES (Li et al., 2024c)	LVLM	Image + Text	Adversarial Prompt Injection	Exploits visual-text vulnerabilities
FigStep (Gong et al., 2023b)	LVLM	Image	Typographic Image Prompting	Bypasses textual filters via typographic images
TypoD (Cheng et al., 2025)	LVLM	Image + Text	Typographic Benchmark	Evaluates typographic image attacks against LVLMs

Table 4: Summary of adversarial attacks targeting VLPs.

4.1 Adversarial Attacks on VLFP

We first discuss adversarial attacks on VLFPs. Such attacks generate adversarial examples by manipulating textual inputs, visual inputs, or both modalities simultaneously. Previous research has primarily concentrated on exploring single-modality and cross-modal perturbation strategies across task-specific classification and information retrieval problems.

Fusion Attack. To exploit vulnerabilities of VLFPs, one notable endeavor is to extend single-source modality adversarial attacks to multiple modalities. Building on the aforementioned insights, Co-Attack (Zhang et al., 2022) proposes a multimodal embedding attack combining image and text perturbations. This attack method integrates PGD iterations with budget constraints for images and BERT-attack for word-level token replacement, effectively perturbing both modalities in information retrieval systems. Unlike Co-Attack, which perturbs text first and then images, Wang et al. (2024g) combine cross-modal adversarial loss, soft-constrained adversarial text loss, and contrastive loss to jointly optimize image and text perturbations to strengthen attack transferability. Meanwhile, VLATTACK (Yin et al., 2023) disrupts image-text representations by enlarging the distance between original and perturbed image features. If this initial image attack fails to alter the model's predictions, the method resorts to BERT-Attack as a word-level replacement attack. Otherwise, it iteratively refines the image-text pair by considering interactions between visual and textual perturbations. SGA (Lu et al., 2023) empirically studies adversarial transferability through set-level guidance data augmentation for expanding multimodal input spaces across different VLFP models and utilizes two-modality interactions as supervision to guide adversarial pair generation. TMM (Wang et al., 2024a) introduces attention-directed feature perturbation to disrupt image and text features, then enhances

adversarial transferability by exploiting text-image alignment within critical attention regions. Gao et al. (2025) propose diversification for the intersection region of the adversarial trajectory and incorporates text-guided adversarial example selection to better exploit cross-modal interactions. The method also mitigates overfitting by steering adversarial text away from the final intersection region during optimization.

Other Attack. OT-Attack (Han et al., 2023) performs optimal transport to map features of image-text pairs using a mutual similarity cost matrix. This approach improves the quality of adversarial examples, aiming to enhance the transferability of adversarial examples. SA-Attack (He et al., 2023) presents a self-augmentation method that increases input diversity by using different augmentation strategies for vision and language modalities to generate adversarial examples. The approach creates adversarial intermediate images and text, utilizing collaborative multimodal interactions to improve adversarial transferability. Shirnin et al. (2024) evaluate the adversarial robustness of VLFP models by utilizing five image perturbation techniques and nine text perturbation techniques to assess performance through comparative analysis. They further explore targeted predictions in relation to category analysis, spurious correlations and aligned modalities.

4.2 Adversarial Attacks on VLCL

VLCL models are designed to learn joint visual-and-text intricate semantic features, enabling remarkable zero-shot learning capabilities. However, VLCLs are susceptible to adversarial attacks that focus on particular similarity and backdoor attacks.

Similarity Attack. AdvCLIP (Zhou et al., 2023b) employs Generative Adversarial Networks (GAN) to create adversarial patches under topology-deviation constraints, minimizing cross-modal embedding similarity to perturb natural images for non-targeted attacks. In addition, ETU (Zhang et al., 2024b) generates universal adversarial perturbations by maximizing embedding distance between matched image-text pairs to enhance data augmentation while preserving semantic coherence, enabling transferability to downstream tasks. Similarly, FGA (Zheng et al., 2024a) is designed to disrupt cross-modal interactions by maximizing the distance between adversarial and original embeddings, targeting alignment mechanisms of VLCL models. Zheng et al. (2024b) propose generating multimodal universal perturbations by exploiting decision boundaries between visual and textual modalities in CLIP. The method supports the creation of global perturbations or adversarial patches, effectively degrading retrieval performance on a variety of tasks. CLIPMasterPrints (Zheng et al., 2024a) introduces a class of adversarial images that are optimized to simultaneously achieve high similarity scores with multiple textual prompts. The attack exploits the modality gap from the misalignment between image and text embedding spaces in contrastive multimodal models. It effectively creates a single universal image that can fool CLIP across many categories, using gradient-based optimization in white-box settings and latent variable evolution in black-box settings.

Backdoor Attack. Carlini et al. (Carlini & Terzis, 2022) evaluate the effectiveness of backdoor patch attacks using a smaller poisoned training dataset against multimodal contrastive pretraining in CLIP, causing incorrect behavior in both feature extractors and zero-shot classifiers. Ko et al. (2023) leverage a three-stage attack strategy by improving membership inference through utilizing cosine similarity thresholds, data augmentations and weakly supervised learning with non-member data. A specific classifier is trained to identify whether a specific image-text pair is derived from the training dataset. RoCLIP (Yang et al., 2023b) applies trigger patches to poisoned images and pairs them with adversarial captions for the training process. It enables the CLIP to misclassify specific test examples with trigger patches to a targeted adversarial label while ensuring benign example predictions remain unchanged.

4.3 Adversarial Attacks on LVLM

LVLMs are capable of performing various multimodal generative tasks, which are integrated into more complex AI systems. Despite the extensive pretraining, studies have shown that LVLMs are vulnerable to malicious attacks (Shayegani et al., 2023a; Liu et al., 2024b; Vatsa et al., 2024; Jin et al., 2024), posing significant security risks and leading to inaccurate predictions and generations. Understanding LVLM safety alignment is essential for developing secure and robust systems. LVLM attacks can be categorized into three categories: adversarial attacks, jailbreak attacks and backdoor attacks.

Adversarial Attack. LVLMs commonly use CLIP-ViT (Radford et al., 2021) or EVA-CLIP (Sun et al., 2023) as their image encoders, and most prior research has concentrated on attacking the visual modality encoder. Researchers identify prefixes that adversaries construct for adversarial images to emit desired harmful outputs, which maximizes the probability of LVLMs generating at least one adversarial token (Carlini et al., 2023). Cui et al. (2024) investigate the adversarial robustness of LVLMs against visual adversarial perturbations using visual occlusions. They illustrate that incorporating object context information significantly improves model robustness across various tasks and datasets. In their approach, visual perturbations are updated to minimize the language modeling loss while textual perturbations are adjusted to maximize the language modeling loss in targeted settings; in non-targeted settings, these objectives are reversed. Schlarmann & Hein (2023) assess the adversarial robustness of target LVLMs through Auto-PGD (APGD) under larger perturbation budgets for untargeted and targeted attacks, suggesting that adding additional context via prompts can enhance the adversarial resilience of LVLMs. VT-Attack (Wang et al., 2024i) is a non-targeted adversarial attack designed to disrupt visual token representations, token relationships, and global semantics within LVLM image encoders. The attack exhibits strong adversarial transferability among LVLMs that share the same image encoder.

Zhao et al. (2023) design transfer-based and query-based attacking strategies to automatically manipulate visual inputs against image-grounded text generation. They demonstrate adversarial transferability across open-source LVLMs. Besides, SparseMA (Yu et al., 2023) is the black-box multimodal attack designed to evaluate the robustness of LVLMs. By introducing sparse perturbations to image patches and textual tokens in a unified discrete space, it simulates adversarial behaviors such as those of illegal merchants and bridges the gap between visual and textual modalities to identify vulnerabilities. Verbose Images (Gao et al., 2024) find an approximately positive linear relationship between energy consumption and latency time with respect to the length of the generated sequence. To exploit this, the method applies imperceptible image perturbations to delay end-of-sequence (EOS) occurrence, enhance output uncertainty, and increase token diversity loss. A temporal weight adjustment algorithm is further introduced to balance these three objectives and maximize sequence length during inference. Zong et al. (2024b) investigate permutation sensitivity in multiple-choice question answering, revealing a critical vulnerability in LLMs and LVLMs. Their study shows that these models are susceptible to adversarial permutations in answer sets, a behavior inconsistent with human invariance to such changes.

Jailbreak Attack. LLMs and LVLMs are increasingly at the forefront of producing content with significant societal implications. However, their growing influence is accompanied by a critical vulnerability: they are highly susceptible to prompt-based adversarial attacks, where minor perturbations in natural language instructions can lead to substantially different predictions. Jailbreak attacks are to specially design prompts that bypass alignment restrictions and elicit harmful responses, generating affirmative toxicity or prohibited responses that exceed the boundaries of safety guardrails, and other containment strategies. This vulnerability arises from LLMs being pre-trained on diverse and large multimodal datasets, which encompass violations, malware content, and harmful information. Wang et al. (2024e) present a comprehensive overview of jailbreaking, focusing on recent advancements in evaluation benchmarks, attack methods and defense strategies. Note that LVLMs remain less explored compared to jailbreaking LLMs. For adversarial manipulation jailbreak, Qi et al. (2024) propose both universal constrained and unconstrained visual and textual adversarial examples to circumvent the safety alignment of LVLMs by maximizing the generation probability using a small corpus of harmful content, forcing the model to execute specific harmful instructions. Additionally, these authors investigate the transferability of the proposed attacks. Furthermore, Universal Master Key (UMK) (Wang et al., 2024c) performs adversarial dual optimizations on both the image prefix and text suffix to effectively jailbreak LVLMs, manipulating them to generate objectionable content. Limitations include generating nonsensical words, exhibiting lower transferability and requiring high computational costs.

For visual injection jailbreak, Jailbreak in Pieces (Shayegani et al., 2023b) investigates a visual injection that generates adversarial images with benign text prompts into the joint embedding space, by embedding various malicious triggers into the joint embedding space. This approach utilizes textual, OCR-textual, and visual triggers to bypass LVLM safeguards and provide unsafe responses. Infectious Jailbreak (Gu et al., 2024) leverages a single LVLM agent to infect almost all other agents to exhibit harmful behaviors without any further intervention. This method employs randomized pair-wise chat for multi-agent interaction and

memory storage to facilitate the jailbreak. Pantazopoulos et al. (2024) investigate the vulnerability of LVLMs for generating harmful content when given a malicous prompt and a semantically relevant image. Their findings reveal that LVLMs employing visual instruction tuning are more susceptible to jailbreak attacks than their base LLMs. Also the study introduces a unified framework for evaluating LVLMs and safety defenses across all stages of training. HADES (Li et al., 2024c) presents a three-stage jailbreak attack framework that hides and amplifies harmful text into crafted images to violate the alignment. It optimizes adversarial images with typographical triggers while maintaining semantic consistency with harmful instructions, thereby bypassing LVLM safety mechanisms. FigStep (Gong et al., 2023b) circumvents safety alignment in LVLMs by converting harmful text prompts into typographic image prompts. This approach exploits the semantic gap between vision and language modalities, allowing harmful instructions embedded in images to evade safeguards and successfully generate prohibited responses.

Backdoor Attack. Backdoor attacks on LVLMs represent a growing threat by exploiting multimodal inputs—inserting triggers into visual and textual modalities that activate malicious behaviors while maintaining normal functionality under benign conditions (Li & Zhang, 2023; Chen & Wang, 2023). BadVLMDriver (Ni et al., 2024) introduces a physical backdoor attack for LVLMs in autonomous driving. BadVLMDriver employs image editing for visual triggers by modifying textual responses using an LLM to inject backdoors into LVLMs. The attack fine-tunes the victim LVLM using both backdoor and benign examples in a white-box manner, ensuring the model behaves correctly under benign inputs while activating attacker-defined behaviors when the visual trigger appears. In both grey-box and black-box settings, Shadowcast (Xu et al., 2024a) manipulates LVLMs through stealthy data poisoning, injecting specifically crafted image-text pairs into the training data. This attack constructs poisoned texts by employing LLMs to paraphrase and modify original captions towards a target concept. Concurrently, it creates poisoned images that visually mimic the target concept while maximizing their feature space distance from the original concept, ensuring stealthiness and effectiveness. AnyDoor (Lu et al., 2024) introduces a test-time backdoor attack targeting LVLMs. Unlike traditional backdoor attacks that rely on contaminated training data, AnyDoor injects backdoors into the textual modality through adversarial test images using universal perturbations.

4.4 Summary of VLP Adversarial Attack

VLPs remain vulnerable to adversarial attacks that exploit imperceptible perturbations to manipulate the original data. These adversarially altered input manipulations can be broadly categorized into poisoning and evasion types: poisoning targets the training sets to corrupt model behaviors, while evasion operates during inference to mislead model predictions. By exploiting these attack, adversaries pose significant threats to VLP systems, targeting intrinsic vulnerabilities inherited from both vision and language encoders. There is no clear understanding of adversarial vulnerabilities of VLPs, pointing to a knowledge gap for further investigations, which remains an active area of research.

5 Adversarial Defense for VLPs

Given the increasing threat of these adversarial manipulations, developing robust defense mechanisms to VLPs is imperative. This is particularly critical when deploying VLPs in sensitive domains where undetected adversarial inputs lead to harmful or misleading outputs. To mitigate these threats, adversarial defenses for VLPs have been proposed and have attracted growing interest (Zhang et al., 2021b), highlighting the importance of developing mechanisms that reduce dependence on non-robust features. Therefore, it is vital to assess their effectiveness and limitations against the emerging adversarial attacks in VLPs. Table 5 summarizes the taxonomy of adversarial defense strategies in VLPs. While these methods demonstrate promising directions, there remains a notable lack of studies specifically targeting adversarial defense for VLFPs.

5.1 VLCL Adversarial Defense

VLCL models require specialized defense approaches due to their unique architectures and training objectives. Recent research has proposed innovative defense methods spanning adversarial contrastive tuning,

Defense Method	Target	Modality	Strategy	Key Contribution
Model Type: VLCL (Vision-Language Contrastive Learning)				
ACL (Jiang et al., 2020)	VLCL	Image + Text	Adversarial Contrastive Training	Enhances representation consistency under perturbations
ADVCL (Fan et al., 2021)	VLCL	Image + Text	Contrastive Learning + Augmentation	Improves robust accuracy in clean and adversarial scenarios
Sim-CLIP (Hossain & Imteaj, 2024)	VLCL	Image + Text	Siamese Architecture + Cosine Loss	Tailors adversarial robustness for vision encoder
MMcOA (Zhou et al., 2024a)	VLCL	Image + Text	Cross-modal Contrastive Loss	Aligns adversarial and clean modalities for robustness
PMG-AFT (Wang et al., 2024d)	VLCL	Image + Text	Auxiliary Branch Alignment	Preserves pretraining features during adversarial fine-tuning
TeCoA (Mao et al., 2023)	VLCL	Image + Text	Text-guided Contrastive Training	Aligns text embeddings with adversarial visual features
LAAT (Li et al., 2024b)	VLCL	Image + Text	Anchor-Based Alignment Loss	Mitigates high cosine similarity issues for robust training
CLAP (Cai et al., 2025)	VLCL	Image + Text	Generative + Contrastive Learning	Enhances adversarial robustness with text augmentation
CleanCLIP (Bansal et al., 2023)	VLCL	Image + Text	Self-Supervised Fine-Tuning	Protects CLIP from backdoor attacks
TGA-ZSR (Yu et al., 2024)	VLCL	Image + Text	Attention Refinement + Constraint	Aligns text-guided attention for adversarial robustness
AdvPT (Zhang et al., 2023)	VLCL	Image + Text	Textual Prompt Optimization	Improves robustness without modifying model architecture
APT (Li et al., 2024a)	VLCL	Image + Text	Robust Prompt Tuning	Optimizes prompts for adversarial resilience
FAP (Zhou et al., 2024b)	VLCL	Image + Text	Cross-modal Prompt Balancing	Balances feature space consistency under adversarial perturbations
TAPT (Wang et al., 2024f)	VLCL	Image + Text	Alignment Loss Optimization	Enhances inference-time robustness
Verma et al. (Verma et al., 2023)	VLCL	Image + Text	Multimodal Contrastive + SSL	Tailors pretraining for backdoor defense
Nirala et al. (Nirala et al., 2024)	VLCL	Image + Text	Open Vocabulary Certification (OVC)	Efficient robustness verification with caching
PromptSmooth. (Hussein et al., 2024)	VLCL	Image + Text	Certified Prompt Learning	Improves certified robustness under Gaussian noise
Model Type: LVLM (Large Vision-Language Models)				
Bhagwatkar et al. (2024b)	LVLM	Image + Text	Prompt Formulation	Enhances robustness via simple prompt modifications
Bhagwatkar et al. (2024a)	LVLM	Image + Text	Vision Encoder + Prompt Design	Evaluates design choices for adversarial robustness
PIP (Zhang et al., 2024e)	LVLM	Image + Text	Attention Pattern Detection	Detects adversarial examples via lightweight classifiers
Bethany et al. (2024)	LVLM	Image + Text	Counterfactual Obfuscation	Explains and masks harmful image regions
Khan and Fu (Khan & Fu, 2024)	LVLM	Image + Text	Consistency-based Detection	Detects unreliable responses from perturbed inputs
FARE-CLIP (Schlarmann et al., 2024)	LVLM	Image + Text	Adversarial Fine-Tuning	Aligns perturbed embeddings with clean embeddings
Qi et al. (2024)	LVLM	Image + Text	Diffusion-based Purification	Mitigates visual adversarial attacks
Zong et al. (2024a)	LVLM	Image + Text	Safety Fine-Tuning	Improves alignment against black-box attacks
Xu et al. (2024b)	LVLM	Image + Text	Cross-modal Semantic Distance	Detects malicious inputs with pre-detection modules
MLLM-Protector (Pi et al., 2024)	LVLM	Image + Text	Binary Detection + Triplet Fine-Tuning	Converts harmful responses into safe ones
Adashield (Wang et al., 2024h)	LVLM	Image + Text	Defense Prompt Optimization	Adheres to safety rules during dialogue interactions
ECSO (Gou et al., 2025)	LVLM	Image + Text	Text Conversion for Safety	Uses LLM safety mechanisms for discrimination
Zhao et al. (2025)	LVLM	Image + Text	Logit Distribution Analysis	Identifies jailbreak attacks through logits
Dress (Chen et al., 2024b)	LVLM	Image + Text	Natural Language Feedback	Aligns safety via critique and refinement
InferAligner (Wang et al., 2024b)	LVLM	Image + Text	Plug-and-Play Safety Vector	Guides alignment with cross-model harmlessness vectors
PSA-VLM (Liu et al., 2024g)	LVLM	Image + Text	Two-Stage Safety Alignment	Aligns visual and textual modalities via projectors and fine-tuning
Chakraborty et al. (Chakraborty et al., 2024)	LVLM	Image + Text	Multimodal Unlearning	Reduces cross-modality harmful generation
Lyu et al. (2024)	LVLM	Image + Text	Poisoned Data Removal	Enhances robustness against training-time backdoors
TIJO (Sur et al., 2023)	LVLM	Image + Text	Trigger Reconstruction	Joint reverse-engineering for image/text triggers
Semantic Shield (Ishmam & Thomas, 2024)	LVLM	Image + Text	External Knowledge Attention	Prevents spurious correlations via contrastive alignment
BaThe (Chen et al., 2024c)	LVLM	Image + Text	Virtual Rejection Embedding	Counters jailbreak backdoors with "wedge" prompt

Table 5: Summary of adversarial defense strategies for VLP

adversarial prompt tuning, backdoor defense and certified robustness. Adversarial contrastive tuning focuses on model fine-tuning, which improves adaptability and preserves generalization capabilities, while adversarial prompt tuning enhances semantic versatility to effectively counteract adversarial examples. Backdoor defense safeguards against poisoned training data, while certified robustness offers provable performance guarantees under bounded perturbations. These approaches underscore the critical need for robust defenses tailored to the specific characteristics of VLCL models.

Adversarial Contrastive Tuning. Adversarial Contrastive Learning (ACL) (Jiang et al., 2020) integrates contrastive learning and adversarial training to enhance adversarial robustness. It leverages perturbed samples during fine-tuning by enforcing representation consistency under data augmentation while maintaining generalization. ADVCL (Fan et al., 2021) combines adversarial examples and high-frequency component augmentations for multi-view contrastive learning with pseudo-label generation, which achieves robust accuracy comparable to standard accuracy under both clean and adversarial scenarios. Sim-CLIP (Hossain & Imteaj, 2024) uses a siamese architecture and employs a stop-gradient mechanism to tailor cosine similarity loss for unsupervised adversarial fine-tuning. It improves the adversarial robustness of CLIP's vision encoder, which can be integrated into existing LVLMs as a frozen vision encoder. MMcOA (Zhou et al., 2024a) considers cross-modal adversarial contrastive losses to align one adversarial modality with the other clean modality and vice versa, fostering robust multimodal representations to defend against multiple adversarial attacks. PMG-AFT (Wang et al., 2024d) introduces an auxiliary branch to preserve the generalization features of pre-trained CLIP. Unlike conventional adversarial fine-tuning, this is achieved by minimizing the feature distance between adversarial examples in the fine-tuned CLIP model and those in the original CLIP.

For the text encoder, TeCoA (Mao et al., 2023) identifies adaptation methods as a key factor in text-guided contrastive adversarial training. The proposed loss aligns text embeddings with adversarial visual features using a small training dataset and can be applied to both fine-tuning and visual prompt tuning. Experimental results indicate that fine-tuning benefits more from textual guidance, whereas prompt tuning shows stronger performance in the absence of textual supervision.

LAAT (Li et al., 2024b) leverages text encoder features for anchor-based adversarial image training with alignment CrossEntropy and Smoothness losses as the optimization objective. This addresses the high cosine similarity problem by increasing the distances of visual features to other anchors while keeping them as close to ground truth anchors as possible. For text augmentation, ACL with Augmented Prompts (CLAP) (Cai et al., 2025) innovatively combines causal generative modeling and contrastive learning with image and text augmentations to disentangle content and style in multimodal representations. This approach enhances the representation learning capabilities of CLIP. In particular, text augmentation demonstrates strong performance in zero-shot and few-shot tasks while also improving robustness against adversarial perturbations. TGA-ZSR (Yu et al., 2024) observes that adversarial perturbations induce shifts in text-guided attention. This approach incorporates two modules: Attention Refinement and Attention-based Model Constraint. The Attention Refinement module aligns the text-guided attention of adversarial examples from the target model with clean examples from the original model. The Attention-based Model Constraint module further constrains the attention distributions of the target and original models on clean examples, improving robustness while maintaining clean performance. Rectify Adversarial Noise (RAN) (Han et al., 2024) constructs medical adversarial noisy datasets using image–caption pairs through multimodal adversarial attacks. It enhances adversarial robustness by fine-tuning vision-language models with covariance loss, consistency loss, and adversarial loss in downstream tasks.

Adversarial Prompt Tuning. These adversarial contrastive tuning approaches are all training-time defense strategies that require pre-tuning for specific downstream tasks. Yet, they are often computationally expensive and difficult to implement on large-scale datasets. In contrast, adversarial prompt tuning approaches present practical solutions for VLCL models, requiring no model retraining or architectural modifications. Adversarial Prompt Tuning (APT) (Li et al., 2024a) involves further fine-tuning of CLIP using a robust text prompt optimization approach to train prompt contexts while keeping image and text encoders frozen. This adversarial prompt tuning method efficiently optimizes prompt contexts while keeping the encoders frozen, improving both adversarial robustness and clean performance. AdvPT (Zhang et al., 2023) introduces optimizing learnable textual vectors for prompt tuning to defend against adversarial images. This prompt tuning defense aligns clean text embeddings with adversarial image embeddings, improving adversarial robustness compared to the vanilla CLIP. Few-shot Adversarial Prompt learning (FAP) (Zhou et al., 2024b) constructs cross-modal prompts under learnable adversarial text supervision to further balance adversarial consistency in the original and adversarial text-image joint feature space. Test-Time Adversarial Prompt Tuning (TAPT) (Wang et al., 2024f) employs alignment loss to optimize the prompt for a given test sample using pre-computed adversarial-clean ImageNet embeddings and multi-view entropy loss to ensure consistent averaged predictions, strengthening the safeguard during inference time.

Backdoor Defense and Certified Robustness. Verma et al. (2023) focus on multimodal contrastive learning and self-supervised learning for mitigating backdoor attacks, highlighting the limitations of current defense methods against backdoor attacks. They emphasize tailored strategies based on specific pre-training objectives even if slightly reducing standard accuracy. CleanCLIP (Bansal et al., 2023) exploits a fine-tuning framework to protect CLIP from various backdoor attacks. It independently relearns unimodal representations through self-supervised learning with modality-specific augmentations and cross-modal contrastive learning. Nirala et al. (2024) propose Open Vocabulary Certification (OVC) for zero-shot CLIP classifiers, which can rapidly identify the most similar certified prompts when the prediction difference falls within a predefined threshold. This method uses a caching mechanism that considerably reduces computation time compared to traditional randomized smoothing. PromptSmooth (Hussein et al., 2024) extends prompt learning to efficiently achieve certified robustness for Med-CLIP under Gaussian noise without retraining. The approach supports zero-shot and few-shot settings, optimizing textual prompts to balance accuracy and robustness while reducing computation costs.

5.2 LVLM Adversarial Defense

To examine the safety of LVLMs, defense measures highlight the importance of preventing unsafe outputs and mitigating potential legal risks. These defense methods can be viewed as countermeasures to strengthen current alignment or design robust alignment mechanisms.

Adversarial Defense. Bhagwatkar et al. (2024b) summarize the impact of model design choices on the adversarial robustness of LVLMs, focusing on various adversarial perturbation techniques. Their findings suggest that prompt formulation without additional context from image or text can effectively enhance robustness against adversarial visual attacks. In subsequent work, the same authors (Bhagwatkar et al., 2024a) provide empirical validation of design principles for incorporating adversarial robustness into LVLMs. Their results demonstrate that targeted modifications to vision encoders and prompt engineering can partially improve adversarial robustness, while increasing image resolution and scaling language model size do not reliably guarantee adversarial robustness. Zhang et al. (2024e) find that visual adversarial and clean examples produce distinct attention patterns in the first generated token under irrelevant probe questions. This method employs a lightweight Support Vector Machine (SVM) classifier to distinguish these attention patterns, providing model-agnostic adversarial detection that operates independently of specific answer categories. Khan & Fu (2024) identify unreliable responses by leveraging the consistency of model outputs across a neighborhood of perturbed visual questions. Since directly sampling neighbors in the feature space is infeasible for black-box models, this approach employs a smaller proxy model to approximate the sampling process. Bethany et al. (2024) investigate the identification of undesirable content by designing a conditional vision LLM to provide explicit reasoning for classifying images as safe or unsafe. The focus is on a counterfactual explanation technique for adaptive segmentation, which precisely and minimally obfuscates sub-object regions in harmful images, providing explanations and targeted obfuscation for safer content moderation. Fine-tuning for FARE-CLIP (Schlarmann et al., 2024) enhances CLIP-ViT robustness in LVLMs through unsupervised adversarial fine-tuning of the vision encoder, using L_2 loss minimization to encourage perturbed image embeddings to remain close to their original representations in the embedding space.

Jailbreak Defense. Qi et al. (2024) adapt purification methods by using Stable Diffusion to defend against visual adversarial examples on targeting LVLMs through image purification. Zong et al. (2024a) curate a safety instruction dataset for standard fine-tuning and post-hoc fine-tuning to ensure safety alignment. The proposed effective fine-tuning framework improves adversarial robustness against black-box attacks and resistance to unsafe instructions without compromising helpfulness. Xu et al. (2024b) propose a plug-and-play pre-detection module that identifies adversarially perturbed images for LVLMs to refuse response generation. They show that the semantic distance between a malicious query and an adversarial image is smaller than that between the same query and a benign image. This cross-modal similarity defense demonstrates strong transferability to both white-box and black-box attacks while offering reduced computational costs and enabling timely inference. MLLM-Protector (Pi et al., 2024) creates the Safe-Harm-10K dataset by using ChatGPT to train the harm binary detector to identify harmless and harmful responses. It then fine-tunes a pretrained LLM using triplets consisting of a text query, a harmful response and a harmless response to convert harmful responses into safe alternatives without compromising performance. Adashield (Wang et al., 2024h) generates diverse defense prompts that adhere to safety rules, avoiding the need for fine-tuning LVLMs or training additional detectors. This framework automatically optimizes defense prompts through interactions between a defender LLM and the target LVLM. Eyes Closed and Safety On (ECSO) (Gou et al., 2025) presents a training-free protection strategy that converts unsafe images into text representations, leveraging the intrinsic safety mechanisms of LLMs for safety discrimination and safe content generation. Zhao et al. (2025) train a linear classifier to analyze the logit distributions of LVLMs, utilizing the first logit distribution with hidden knowledge to identify jailbreak attacks.

Safety Alignment. Another LVLM safety research direction focuses on analyzing safety alignment (Dai et al., 2023). The method proposed in (Chen et al., 2024b) effectively learns from natural language feedback using critique and refinement techniques that are non-differentiable, applying them to trained LVLMs. This approach improves alignment with human preferences and multi-turn interaction capabilities, reducing harmful responses while increasing helpfulness and honesty. InferAligner (Wang et al., 2024b) achieves harmlessness alignment at inference time without requiring additional training. This plug-and-play method employs safety steering vectors extracted from aligned models to provide cross-model guidance, guiding target models toward harmless behavior. Progressive Safety Alignment for VLMs (PSA-VLM) (Liu et al., 2024g) proposes a two-stage training framework: first, aligning visual modality safety modules through safety projectors, tokens and heads; then, fine-tuning the LLM backbone to learn safety-aligned features in the second stage. These advancements demonstrate the capability of LVLMs to enhance model confidence in identifying malicious content. Chakraborty et al. (2024) compare text-only and multimodal unlearning

approaches to enforce cross-modality safety alignment in LVLMs, aiming to prevent harmful content generation. Textual unlearning not only significantly reduces attack success rates but also achieves remarkable harmlessness against cross-modality attacks. Although the aforementioned progress demonstrates advances in safety alignment, persistent efforts remain crucial to address alignment vulnerabilities and enhance safety and reliability in this field.

Backdoor Defense. With the growing threat of data poisoning, recent studies have explored innovative defense mechanisms (Lyu et al., 2024). Designing effective backdoor defense strategies tailored to diverse pre-training paradigms is critical for detecting and eliminating poisoned data. Trigger Inversion using Joint Optimization (TIJO) (Sur et al., 2023) proposes jointly reverse-engineering triggers in both image and text modalities. Unlike existing methods that often focus on isolated modalities, TIJO optimizes within the object detection bounding box feature space for images and employs universal adversarial triggers for text. The innovation stems from the integrated approach to the image-text pipeline by reconstructing visual triggers within the feature space of detected object bounding boxes. Semantic Shield (Ishmam & Thomas, 2024) proposes a contrastive training framework for LVLMs to mitigate backdoor and poisoning attacks. This approach utilizes external knowledge from a language model to constrain attention mechanisms, ensuring that the model focuses on visual regions that are semantically aligned with external knowledge. By enforcing this alignment, the method prevents the model from learning spurious correlations introduced by poisoned data. Backdoor Trigger Shield (BaThe) (Chen et al., 2024c) addresses the continuous nature of image signals that enable the direct injection of harmful intentions. In contrast to text-based LLMs, where adversaries use discrete token algorithms to conceal malicious content, BaThe embeds a virtual rejection prompt termed a "wedge" into soft text embeddings to trigger rejection responses during training, thereby mitigating harmful instructions.

5.3 Summary of VLP Adversarial Defense

In this section, we review existing defense strategies for VLPs, which has gained profound insights into the development of robust defense mechanisms. Despite this progress, significant challenges remain in achieving practical and effective defenses. Armed with these findings, future research can focus on strengthening cross-modal alignments and developing robust defense mechanisms against unlabeled data.

6 Adversarial Attack Impact on VLP Applications

We note that vision-language model pre-training enables models to be applied directly to downstream tasks or adapted via fine-tuning for task-specific requirements. This section reviews these downstream tasks and analyzes the adversarial vulnerabilities of VLPs. These vulnerabilities raise critical concerns for safety-sensitive multimodal applications, highlighting the necessity of exploring adversarial robustness to ensure reliable real-world deployment in real-world scenarios.

6.1 Multimodal Classification

VQA (Antol et al., 2015) represents machine understanding of images by answering questions from an open-ended answer set. It enables question-answering systems to comprehend visual content for a given image and question, providing accurate natural language responses. The goal of VQA is to select the most accurate answer from a candidate answer list. To address VQA's shortcomings, GQA (Hudson & Manning, 2019) introduces a new dataset designed to mitigate biased reasoning. The GQA dataset provides visual graphs, images, and questions for visual scene reasoning and compositional question answering. GQA leverages visual scene understanding capabilities to substantially reduce biases. Visual Commonsense Reasoning (VCR) (Zellers et al., 2019) extends beyond object recognition by focusing on holistic visual understanding and inferring complex object relationships to achieve higher-order cognitive reasoning. It aims to provide correct answers with rationale justification from given image-question pairs. The key issue is to convert rich annotations into multiple-choice questions with minimal prejudice. Natural Language for Visual Reasoning (NLVR) (Suhr et al., 2017) is a visual-language reasoning dataset that focuses on joint reasoning over two images and one natural language description. This design promotes compositional semantics through complex

visual reasoning to determine binary predictions of whether the description accurately describes the visual inputs. Visual Dialogue (VD) (Das et al., 2017) explicitly incorporates dialogue history, engaging in a series of questions about visual content. It generates responses from candidate option lists, requiring sophisticated solutions to maintain multi-turn conversations with users in natural language while grounding information in images. Visual Grounding (VG) (Yu et al., 2018) localizes corresponding objects or regions in images given natural language expressions. VG aims to comprehend complex reasoning within natural language expressions, such as phrases, sentences, or multi-turn dialogues. The VG task focuses on identifying the most relevant spatial and semantic relationships among target objects in images. Visual Entailment (VE) (Xie et al., 2019) is a multimodal reasoning task that addresses visual intelligence in real-world settings. In this task, an image serves as the premise, while a natural language sentence acts as a hypothesis describing the image. The goal of VE is to enable systems to perform logical inference and determine whether the hypothesis is entailed by the visual content.

Existing VLP adversarial example generation methods are typically developed and evaluated on specific datasets, with limited exploration of adversarial transferability across diverse downstream tasks. To develop downstream-agnostic adversarial attacks, it is essential to verify their effectiveness across a wide range of tasks using diverse VLP models. For example, ensuring consistent performance under adversarial conditions is still an open problem, as it remains unclear whether VQA adversarial examples can generalize to VG tasks. Furthermore, in real-world applications such as autonomous driving, medical image diagnosis and sensitive financial predictions, understanding and mitigating potential adversarial vulnerabilities becomes paramount. These high-stakes scenarios underscore the critical need to improve adversarial robustness and develop reliable verification frameworks.

6.2 Multimodal Generation

Multimodal Text Generation (MTG) (Cho et al., 2021; Lin et al., 2021) focuses on aligning different modalities, such as text and images for text generation. It reflects the capability of language models to interpret visual and textual cross-modal information to produce coherent natural language descriptions. Multimodal Machine Translation (MMT) (Specia et al., 2016) involves generating target language sentences from source language descriptions while incorporating additional information from corresponding images. This task extends traditional machine translation by leveraging visual cues to improve translation quality and disambiguate textual content. Image Captioning (Hossain et al., 2019; Lin et al., 2014) aims to generate descriptive natural language sentences for given images. This task requires understanding visual scenes, object attributes, and their relationships to produce semantically accurate textual descriptions. Novel Object Captioning at Scale (NoCaps) (Agrawal et al., 2019) is a large-scale benchmark designed to evaluate image captioning models on novel object recognition. It measures the ability to generalize to unseen visual concepts and describe objects that do not appear in the caption training data. Image Text Retrieval (ITR) (Cao et al., 2022) is designed for retrieving or generating related information from source modalities and mapping it to heterogeneous target modalities. It encompasses image-to-text, image-to-image, and text-to-image perspectives. Visual Linguistic Navigation (VLN) (Anderson et al., 2018) involves communication between an oracle (human) and an agent (robot) that can interpret natural language navigation instructions. The agent interprets these instructions to navigate and find optimal paths between states in various environments. This human-robot interaction links natural language to visual information, enabling autonomous action selection for subsequent states.

Adversarial attacks against VLPs pose significant threats by disrupting multimodal generation capabilities and circumventing built-in safety mechanisms. These adversarial attacks against LVLMs manifest both in digital environments and in physical deployments, leading to unintended model behaviors in applications ranging from question-answering systems for chatbots and embodied AI for robotic systems. In particular, jailbreak attacks to LVLMs represent a critical category that deliberately bypass safety alignments to elicit harmful responses. Additionally, many safety-critical applications still lack robust and publicly available evaluation datasets. VLPs can inherit biases from their training data sources, which can ultimately affect prediction accuracy and fairness. Moreover, the risk of incorporating private information, such as phone numbers and email addresses, into LLM training datasets raises concerns, as privacy information leakage poses significant threats to public safety and trust. As LVLMs become increasingly integrated into real-world

AI-driven systems, it is imperative to evaluate their adversarial robustness across diverse tasks to ensure safe and responsible deployment.

6.3 Summary of VLP Datasets

We categorize the datasets of VLPs into two types: classification and generation tasks. VLPs rely on multimodal data that integrates visual and linguistic modalities to enhance prediction accuracy. However, they exhibit varying degrees of adversarial vulnerability across complex downstream tasks. Adversarial attack research has extended into multimodal settings to investigate the adversarial robustness of VLPs, enabling the adaptation of conventional methods to the challenges of multimodal attack research.

7 Future Research Directions and Open Issues

Multimodal adversarial robustness has a shorter research history compared to single-modality approaches, yet multimodal vulnerabilities pose significant security threats to complex AI systems. With the increasing application of VLPs such as GPT-4 (Achiam et al., 2023), multimodal robust learning has become a critical aspect in the development stage. Developing trustworthy VLPs remains a major challenge and an open research question, as ensuring trustworthiness requires a thorough understanding of their vulnerabilities even when these models exhibit excellent performance. The adversarial robustness of VLPs in real-world deployments remains largely unclear and is still in its infancy, necessitating urgent research—particularly given the complex challenges of guaranteeing robustness across diverse scenarios. In response, this review elaborates on open issues and key challenges while proposing potential research directions. The adversarial robustness of VLPs in real-world deployments remains largely unclear and is still in its infancy, necessitating urgent research in this area. In response, this review elaborates on open issues and key challenges while proposing potential research directions.

7.1 Multimodal Learning

Model Architecture. The fundamental architecture plays a more crucial role in adversarial robustness than model size (Su et al., 2018). Existing research on VLPs focuses on novel architectures and downstream applications, which remain insufficiently explored in the context of robust learning under adversarial perturbations. This gap offers insights for understanding model structural characteristics and conducting architecture-level analyses that can enhance adversarial robustness. Given the flexibility of transformer architectures in joint representation learning, enhancing adversarial robustness from an architectural perspective involves incorporating specialized modules, such as robust visual and textual encoders before integrating for outputs. Alternatively, a prominent future research direction is VLP architectural modifications that reduce the reliance on task-specific designs while maintaining model capacity for effective cross-modal interaction. Such architectural changes can boost VLP performance and enhance adversarial robustness.

Model Interpretability. Robust VLPs should not only exhibit high performance but also offer interpretability—the ability of humans to comprehend model decision-making processes. Interpretable VLPs allow researchers to understand how each modality shapes decisions through cross-modal interactions (Chefer et al., 2021). Ultimately, such interpretability and transparency enable effective collaboration between users and models, allowing users to accept or reject predictions based on informed judgment. By exposing the VLPs, the study on tree-augmented vision-language (3VL) (Yellinek et al., 2023) illustrates that integrating tree structures with anchor inference and differential relevance can improve VLP interpretability for model behaviors. This strategy aims to develop VLPs that are both robust and interpretable, improving the reliability in security-sensitive applications

Causal Inference. Causal inference represents a promising research direction in multimodal intelligence. It provides interpretable explanations for predictions to increase trustworthiness and enable effective human-AI interaction. In multimodal intelligence, causal inference aims to explore relationships between cause and effect, how different modalities influence each other, and how integrated information from both modalities contributes to decision-making in VLPs. To assess causal reasoning fairly and accurately, CELLO (Chen et al., 2024a) introduces a benchmark dataset to comprehensively examine causality capabilities in LVLMs

across four aspects: discovery, association, intervention, and counterfactual reasoning. Experimental investigations reveal notable limitations in the causal capabilities of LVLMs, while also demonstrating improved causal performance using the proposed chain-of-thought prompting strategy. Multimodal causality can contribute to robust learning in VLPs by distinguishing between genuine and adversarial examples through causal relationships.

7.2 Multimodal Attacks

Adversarial Tactics. VLPs employ attention mechanisms across modalities over the sequence dimension. Modal entities are mutually interconnected through dot-product similarity, enabling each element to attend to others across modalities. Building upon this theoretical foundation, VLP fusion strategies aim to provide multimodal contextualized representations, thereby enhancing cross-modal learning techniques (Liang et al., 2022). Multimodal attacks that simultaneously target images and text have been shown to be more effective than on a single modality attacks. These adversarial tactics exploit the alignment and fusion to disrupt the semantic correspondence between modalities. Understanding and mitigating these multimodal attacks by crafting cross-modal perturbations can substantially contribute to the development of more security multimodal systems.

Adversarial Transferability. Adversarial transferability refers to the phenomenon where adversarial examples generated from a source model can effectively misclassify different unseen target models. For VLPs, this problem entails investigating transferable adversarial attacks across the VLP family and identifying factors that contribute to attack transferability, which has received limited attention in existing research. Luo et al. (2024a) address this issue by simultaneously optimizing image inputs and multiple prompts in opposite directions while updating image perturbations to achieve cross-prompt adversarial transferability. Gao et al. (2025) further propose a diversification strategy for the intersection regions along the adversarial trajectory, which aims to increase transferability across various VLPs. These suggested directions can help improve adversarial robustness against evolving adversarial tactics and develop more generalizable defense strategies.

Benchmarks. Numerous safety-related benchmarks have been proposed to evaluate adversarial robustness in LVLMs. HADES (Li et al., 2024c) contributes a dataset comprising 750 harmful text-image pairs distributed across five distinct scenarios, using ChatGPT for instruction modification and diffusion models for harmful image generation. SafeBench (Gong et al., 2023b) creates a safety benchmark covering 10 safety topics with 500 harmful queries and images. MM-SafetyBench (Liu et al., 2024f) investigates how LVLMs perform under query-relevant images with malicious key phrases transformed into images comprising typography approaches. This benchmark proposes effective safety prompt strategies to enhance LVLM resilience and includes 13 scenarios containing 5,040 text-image pairs for image-based jailbreak evaluation, assessing 12 state-of-the-art LVLMs by those equipped with safety alignment mechanisms. JailBreakV-28K (Luo et al., 2024b) collects 2,000 harmful queries with 16 safety policy adversarial scenarios and contains 28,000 jailbreak text-image pairs. This dataset is strategically divided into two categories: 20,000 text-based transferability cases of LLM jailbreak techniques to LVLMs and 8,000 image-based LVLM jailbreak attacks, highlighting critical vulnerabilities. Cheng et al. (2025) introduce a typographic benchmark dataset for evaluating the vulnerability of LVLMs to existing typographic attacks, where malicious text embedded in images disrupts vision encoders, emphasizing the need for improving resistance to typographic vulnerabilities.

7.3 Adversarial Defense

Adversarial Training. Adversarial training (Madry et al., 2018; Bai et al., 2021) has received considerable attention and is generally considered one of the most effective adversarial defense strategies, although its effectiveness is limited by the characteristics of specific datasets and tasks. While adversarial training enhances adversarial robustness by augmenting standard training procedures, it negatively impacts generalization performance. Compared to standard training, adversarial training is computationally expensive due to the inner maximization requirement for adversarial loss through applying perturbations in the fusion embedding space. Gan et al. (2020) propose adversarial training to improve performance without incurring additional computational usage by creating adversarial perturbations in the multimodal embedding space.

However, this approach focuses on improving model performance rather than defending against adversarial attacks. One natural extension to adversarial training is to create general-purpose VLP systems that do not rely heavily on task-specific properties, thereby reducing the negative impact on generalization performance without incurring significant theoretical limitations or computational costs.

Certified Robustness. In preliminary explorations, we emphasize that adversarial robustness serves as a meaningful additional comparison point beyond standard evaluations (Carlini et al., 2019). Relying solely on empirical adversarial robustness evaluations is insufficient for reliable deployment in real-world scenarios. One of the key challenges in this area is the lack of certified robustness guarantees. Certified robustness aims to formally ensure that model outputs remain consistent across both clean and perturbed inputs, providing theoretical guarantees of robustness. The theoretical understanding of robust DNNs remains inadequate, and this limitation extends to even more complex VLPs. For instance, Nirala et al. (2024) introduce open vocabulary verification to certify the robustness of CLIP against adversarial attacks using incremental randomized smoothing. This approach leverages novel prompts as perturbations of nearby classifiers in the certification process using caching-based acceleration. In certified robustness research, the primary goal is to develop certification and verification approaches serving as quality assurance in response to perturbed data. Therefore, it is essential to verify VLPs to ensure they deliver acceptable performance under uncertain adversarial conditions.

Adversarial Purification. The diffusion-based generative models has introduced purification methods that are model-agnostic and without the need for task-specific retraining and architecture modifications (Shi et al., 2020; Yoon et al., 2021). Adversarial purification (Wu et al., 2022) is a safeguard that removes adversarial perturbations by generating purified inputs comparable to the original inputs, ensuring correct predictions by victim models. One key advantage of this approach is that it reduces the requirement for deeper understanding of model internals and defends against unseen threats without identifying specific adversarial attack types. Recent work on VLP adversarial purification (Qi et al., 2024) demonstrates implementation in a plug-and-play manner to counter visual adversarial examples. This process purifies visual adversarial examples by reconstructing inputs through generative models that guide them back onto the data manifold, offering a flexible and modular approach that can be integrated into existing VLPs before prediction.

Prompt Tuning. Jailbreak prompts (Deng et al., 2023) are specially designed to bypass service provider constraints by exploiting AI alignment safeguards or other confinement measures. Additionally, prompt injection (Zhang et al., 2023; Kan et al., 2023) focuses on controlling inputs by inserting adversarially constructed prompts. Conversely, prompt techniques can also be adapted for adversarial defense, by crafting robust prompts that resist jailbreak attempts and mitigate prompt injection. Inspired by advancements in prompt engineering, prompt tuning defense is a strategy designed to improve the adversarial robustness of VLPs without retraining models. By optimizing input prompts without altering model architectures or parameters, this methodology effectively mitigates adversarial vulnerabilities and provides a lightweight defense mechanism for existing VLPs. Chen et al. (2023a) propose a class-wise adversarial visual prompting (C-AVP) approach that generates class-specific visual prompts to enhance adversarial robustness. C-AVP benefits from ensemble prompts and optimized interrelationships. This approach overcomes the limitations of universal prompts for robust learning against adversarial perturbations. Devoting research effort to exploring defense strategies such as prompt tuning and adversarial prompt detection is an important future research direction.

7.4 Safety Enhancement

Machine Unlearning. Machine unlearning is an emerging paradigm that removes particular training data from machine learning models without requiring complete retraining. In the context of LVLMs, this capability allows for the efficient elimination of adversarial data while preserving overall model performance (Chakraborty et al., 2024). Therefore, machine unlearning techniques offer a promising approach for mitigating adversarial and jailbreak attacks. Specifically, machine unlearning provides a safeguard mechanism to revoke certain learned capabilities by removing adversarial or sensitive data from the training process. As multimodal machine unlearning remains in its early stages, this field involves understanding how removing sensitive or private information impacts fused multimodal representations while ensuring consistent model performance after unlearning.

Model Editing. Although the pretraining approach has proven to be advantageous for multimodal tasks, a limitation of LVLM pretraining lies in its dependence on large-scale datasets and substantial computational power. This raises the challenge of maintaining accurate and current knowledge without incurring substantial retraining costs. Model editing refers to manipulating the behavior of models in specific domains by efficiently modifying model parameters or integrating auxiliary modules while preserving the original and unaltered models. Recent research in model editing aims to address undesirable outputs while preserving the integrity of interpretations and without degrading performance on unrelated inputs. For example, Geva et al. (2022) introduce model editing to suppress potentially harmful text generation. And MMEdit (Cheng et al., 2023) presents a multi-modality model editing benchmark that achieves high fidelity across various vision-language tasks under three evaluation principles: reliability, locality, and generality. Thus, model editing can potentially mitigate toxic and illegal knowledge stored in LLMs in response to certain prompts, which is vital for enhancing the security and robustness of LVLMs.

Mitigating Hallucination. There is growing concern regarding hallucination defined as the generation of content that is unfaithful to the input. Specifically, LLMs tend to hallucinate undesired text (Ji et al., 2023), and LVLMs generate nonexistent objects in images (Li et al., 2023c). Observe–Reason–Critique–Act (ORCA) (Yu et al., 2025) introduces a training-free agentic reasoning framework that addresses both hallucination mitigation and adversarial robustness in LVLM test time. It leverages lightweight external vision models and performs cross-model consistency verification to identify and correct unreliable predictions. This understanding helps develop more sophisticated hallucination detection and defense mechanisms against adversarial attacks.

8 Conclusion

The advent of vision-language pretraining (VLP) models constitutes a significant milestone in AI, advancing their capacity to comprehend and operate across multiple modalities. Despite their transformative promise, VLPs remain highly susceptible to increasingly sophisticated adversarial threats. In this review, we provide a systematic overview of recent advances in both adversarial attacks and defenses for VLPs. We examine attack methods that expose vulnerabilities in VLPs, and discuss defense strategies aimed at enhancing their security in real-world, high-stakes scenarios. We also explore potential future research directions in this domain, highlighting opportunities for innovation and advancement. Ultimately, we hope this review helps address the multimodal vulnerabilities of VLPs while providing a foundational framework for future research in this rapidly evolving field.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, pp. 8948–8957, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 23716–23736, 2022.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 3674–3683, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. BadCLIP: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24239–24250, 2024.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. CleanCLIP: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 112–123, 2023.

Mazal Bethany, Brandon Wherry, Nishant Vishwamitra, and Peyman Najafirad. Image safeguarding: Reasoning with conditional vision language model and obfuscating unsafe content counterfactually. In *Association for the Advancement of Artificial Intelligence*, 2024.

Rishika Bhagwatkar, Shravan Nayak, Pouya Bashivan, and Irina Rish. Improving adversarial robustness in vision-language models with architecture and prompt design. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 17003–17020, 2024a.

Rishika Bhagwatkar, Shravan Nayak, Reza Bayat, Alexis Roger, Daniel Z Kaplan, Pouya Bashivan, and Irina Rish. Towards adversarially robust vision-language models: Insights from design choices and prompt formatting techniques. In *International Conference on Learning Representations Next Generation of AI Safety Workshop*, 2024b.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *In Proceedings of the 34th International Conference on Neural Information Processing Systems, NeurIPS*, pp. 1877–1901, 2020.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, Rita Cucchiara, et al. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Yichao Cai, Yuhang Liu, Zhen Zhang, and Javen Qinfeng Shi. Clap: Isolating content from style through contrastive learning with augmented prompts. In *European Conference on Computer Vision*, pp. 130–147. Springer, 2025.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. In *In Proceedings of International Joint Conference on Artificial Intelligence, IJCAI*, 2022.

Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *the ACM SIGSAC conference on computer and communications security*, 2019.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2022.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? In *Neural Information Processing Systems*, 2023.

Trishna Chakraborty, Erfan Shayegani, Zikui Cai, Nael Abu-Ghazaleh, M Salman Asif, Yue Dong, Amit Roy-Chowdhury, and Chengyu Song. Can textual unlearning solve cross-modality safety alignment? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9830–9844, 2024.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021.

Aochuan Chen, Peter Lorenz, Yuguang Yao, Pin-Yu Chen, and Sijia Liu. Visual prompting for adversarial robustness. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. VLP: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023b.

Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. Cello: Causal evaluation of large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22353–22374, 2024a.

Xiaoyu Chen and Haoran Wang. Vlmguard: A robust defense against multimodal backdoor attacks in vision-language models. In *Proceedings of the 2023 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4567–4576. IEEE, 2023.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14239–14250, 2024b.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *In Proceedings of the European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

Yulin Chen, Haoran Li, Zihao Zheng, and Yangqiu Song. Bathe: Defense against the jailbreak attack in multimodal large language models by treating harmful instruction as backdoor trigger. *arXiv preprint arXiv:2408.09093*, 2024c.

Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language models. In *European Conference on Computer Vision*, pp. 179–196. Springer, 2025.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Zhang Ningyu. Can we edit multimodal large language models? In *EMNLP*, 2023.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *In Proceedings of the International Conference on Machine Learning, ICML*, pp. 1931–1942. PMLR, 2021.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24625–24634, 2024.

Bo Dai, Lei Yang, Yu Yang, et al. Pixel-bert: Aligning image pixels with text by deep multi-modal pretraining. In *Computer Vision and Pattern Recognition*, 2020.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *Proceedings of the International Conference on Learning Representations*, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *International Conference on Neural Information Processing Systems*, 2024.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, et al. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*, 2024.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 326–335, 2017.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies)*, pp. 4171–4186, 2019.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10995–11005, 2023.

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 *In Proceedings of the International Conference on Learning Representations*, 2021.

Ziyi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *In the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18166–18176, 2022.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. In *In Proceedings of the 35th International Joint Conference on Artificial Intelligence, IJCAI*, pp. 5436–5443, 2022.

Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? In *Proceedings of the Neural Information Processing Systems*, pp. 21480–21492, 2021.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 6616–6628, 2020.

Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. In *Proceedings of the International Conference on Learning Representations*, 2024.

Sensen Gao, Xiaojun Jia, Xuhong Ren, Ivor Tsang, and Qing Guo. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *European Conference on Computer Vision*, pp. 442–460. Springer, 2025.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, 2022.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: cyclic contrastive language-image pretraining. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 6704–6719, 2022.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-GPT: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023a.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*, 2023b.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision*, pp. 388–404. Springer, 2025.

Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. In *The International Conference on Machine Learning*, 2024.

Dongchen Han, Xiaojun Jia, Yang Bai, Jindong Gu, Yang Liu, and Xiaochun Cao. Ot-attack: Enhancing adversarial transferability of vision-language models via optimal transport optimization. *arXiv preprint arXiv:2312.04403*, 2023.

Xu Han, Linghao Jin, Xuezhe Ma, and Xiaofeng Liu. Light-weight fine-tuning method for defending adversarial noise in pre-trained medical vision-language models. In *Findings of the Association for Computational Linguistics*, pp. 10784–10799, 2024.

Bangyan He, Xiaojun Jia, Siyuan Liang, Tianrui Lou, Yang Liu, and Xiaochun Cao. SA-Attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation. *arXiv preprint arXiv:2312.04913*, 2023.

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*, pp. 770–778, 2016.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations, (ICLR)*, 2020.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

Md Zarif Hossain and Ahmed Imteaj. Sim-clip: Unsupervised siamese adversarial fine-tuning for robust and semantically-rich vision-language models. *arXiv preprint arXiv:2407.14971*, 2024.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2024.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Computer Vision and Pattern Recognition*, pp. 12976–12985, 2021.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, CVPR*, pp. 6700–6709, 2019.

Noor Hussein, Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Promptsmooth: Certifying robustness of medical vision-language models via prompt learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 698–708. Springer, 2024.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. *Zenodo*, 2021.

Alvi Md Ishmam and Christopher Thomas. Semantic shield: Defending vision-language models against backdooring and poisoning via fine-grained knowledge alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24820–24830, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. In *Proceedings of the Neural Information Processing Systems*, pp. 16199–16210, 2020.

Haibo Jin, Leyang Hu, Xinuo Li, Peiyan Zhang, Chonghan Chen, Jun Zhuang, and Haohan Wang. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv:2407.01599*, 2024.

Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15670–15680, 2023.

Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10854–10863, 2024.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *In Proceedings of the International Conference on Machine Learning, ICML*, pp. 5583–5594. PMLR, 2021.

Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4871–4881, 2023.

Jie Li and Wei Zhang. Multimodal backdoor attacks and defenses: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1–19, 2023.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *In Proceedings of the Advances in neural information processing systems, NeurIPS*, volume 34, pp. 9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *In Proceedings of the International Conference on Machine Learning, ICML*, pp. 12888–12900. PMLR, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023a.

Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24408–24419, 2024a.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5265–5275, 2020a.

Xiao Li, Wei Zhang, Yining Liu, Zhanhao Hu, Bo Zhang, and Xiaolin Hu. Language-driven anchors for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24686–24695, 2024b.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *In Proceedings of the European Conference on Computer Vision, ECCV*, pp. 121–137. Springer, 2020b.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *In Proceedings of the International Conference on Learning Representations, (ICLR)*, 2022b.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23390–23400, 2023b.

Yen-Chun Li, Weijie Su, Ming-Yu Chang, et al. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2020c.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023c.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *In Proceedings of the European Conference on Computer Vision*, 2024c.

Yue Li, Xiao Li, Hao Wu, Yue Zhang, Xiuzhen Cheng, Sheng Zhong, and Fengyuan Xu. Attention is all you need for llm-based code vulnerability localization. *arXiv preprint arXiv:2410.15288*, 2024d.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision, ECCV*, pp. 740–755. Springer, 2014.

Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 7005–7015, 2021.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*, 2024b.

Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. SPHINX-X: Scaling data and parameters for a family of multi-modal large language models. In *Proceedings of the International Conference on Machine Learning*, 2024c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024d.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Conference on Neural Information Processing Systems*, 2024e.

Haoyang Liu, Maheep Chaudhary, and Haohan Wang. Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives. *arXiv preprint arXiv:2307.16851*, 2023b.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. MM-safetyBench: A benchmark for safety evaluation of multimodal large language models. In *In Proceedings of the European Conference on Computer Vision*, 2024f.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zhendong Liu, Yuanbi Nie, Yingshui Tan, Xiangyu Yue, Qiushi Cui, Chongjun Wang, Xiaoyong Zhu, and Bo Zheng. Enhancing vision-language model safety through progressive concept-bottleneck-driven alignment. *arXiv preprint arXiv:2411.11543*, 2024g.

Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqing Yang. Vision-and-language pretrained models: A survey. In *International Joint Conference on Artificial Intelligence*, pp. 5530–5537, 2022.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 102–111, 2023.

Dong Lu, Tianyu Pang, Chao Du, Qian Liu, Xianjun Yang, and Min Lin. Test-time backdoor attacks on multimodal large language models. *arXiv preprint arXiv:2402.08577*, 2024.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *In Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 13–23, 2019.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. In *Proceedings of the International Conference on Learning Representations*, 2024a.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024b.

Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, pp. 467–483, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *In Proceedings of the 2018 International Conference on Learning Representations (ICLR)*, 2018.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *International Conference on Learning Representations*, 2023.

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, 71:1183–1317, 2021.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision (ECCV)*, pp. 529–544. Springer, 2022.

Katsuaki Nakano, Michael Zuzak, Cory Merkel, and Alexander C Loui. Trustworthy and robust machine learning for multimedia: Challenges and perspectives. In *IEEE International Conference on Multimedia Information Processing and Retrieval*, pp. 522–528, 2024.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *In Proceedings of the 28th International Conference on Machine Learning, ICML*, 2011.

Zhenyang Ni, Rui Ye, Yuxi Wei, Zhen Xiang, Yanfeng Wang, and Siheng Chen. Physical backdoor attack can jeopardize driving with vision-large-language models. In *International Conference on Machine Learning Trustworthy Multi-modal Foundation Models and AI Agents (TiFA) workshop*, 2024.

Ashutosh Nirala, Ameya Joshi, Soumik Sarkar, and Chinmay Hegde. Fast certification of vision-language models using incremental randomized smoothing. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 252–271. IEEE, 2024.

Georgios Pantazopoulos, Amit Parekh, Malvina Nikandrou, and Alessandro Suglia. Learning to see but forgetting to follow: Visual instruction tuning makes llms more prone to jailbreak attacks. In *Safety4ConvAI: The Third Workshop on Safety for Conversational AI LREC-COLING*, pp. 40–51, 2024.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. Mllm-protector: Ensuring mllm's safety without hurting performance. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3677–3685, 2023.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *International Conference on Machine Learning*, 2024.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv e-prints*, pp. arXiv–2310, 2023a.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The International Conference on Learning Representations*, 2023b.

Changhao Shi, Chester Holtz, and Gal Mishne. Online adversarial purification based on self-supervised learning. In *International Conference on Learning Representations (ICML)*, 2020.

Alexander Shirnin, Nikita Andreev, Sofia Potapova, and Ekaterina Artemova. Analyzing the robustness of vision & language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Amanpreet Singh et al. FLAVA: A foundational language and vision alignment model. In *Advances in Neural Information Processing Systems*, 2021.

Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 543–553, 2016.

Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–648, 2018.

Weijie Su, Chunyuan Sun, Zhe Li, Yukun Zhang, et al. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pp. 11–23, 2023.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 217–223, 2017.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-CLIP: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Indranil Sur, Karan Sikka, Matthew Walmer, Kaushik Koneripalli, Anirban Roy, Xiao Lin, Ajay Divakaran, and Susmit Jha. Tijo: Trigger inversion with joint optimization for defending multimodal backdoored models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 165–175, 2023.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michael Tschannen, Basil Mustafa, and Neil Houlsby. CLIPPO: Image-and-language understanding from pixels only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11006–11017, 2023.

Mayank Vatsa, Anubhooti Jain, and Richa Singh. Adventures of trustworthy vision-language models: A survey. In *Association for the Advancement of Artificial Intelligence*, pp. 22650–22658, 2024.

Sahil Verma, Gantavya Bhatt, Avi Schwarzschild, Soumye Singhal, Arnav Mohanty Das, Chirag Shah, John P Dickerson, Pin-Yu Chen, and Jeff Bilmes. Effective backdoor mitigation in vision-language models depends on the pre-training objective. *Transactions on Machine Learning Research*, 2023.

Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *IEEE Symposium on Security and Privacy (SP)*, pp. 102–102, 2024a.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. InferAligner: Inference-time alignment for harmlessness through cross-model guidance. In *the Association for Computational Linguistics*, 2024b.

Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. White-box multimodal jailbreaks against large vision-language models. In *ACM Multimedia*, 2024c.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24502–24511, 2024d.

Siyuan Wang, Zhuohan Long, Zhihao Fan, and Zhongyu Wei. From llms to mllms: Exploring the landscape of multimodal jailbreaking. In *Conference on Empirical Methods in Natural Language Processing*, pp. 17568–17582, 2024e.

Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. *arXiv preprint arXiv:2411.13136*, 2024f.

Youze Wang, Wenbo Hu, Yinpeng Dong, Hanwang Zhang, Hang Su, and Richang Hong. Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning. In *IEEE Transactions on Multimedia*, 2024g.

Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In *European Conference on Computer Vision*, 2024h.

Yubo Wang, Chaohu Liu, Yanqiu Qu, Haoyu Cao, Deqiang Jiang, and Linli Xu. Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1072–1081, 2024i.

Maurice Weber, Xiaojun Xu, Bojan Karlaš, Ce Zhang, and Bo Li. Rab: Provable robustness against backdoor attacks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1311–1328. IEEE, 2023.

Quanlin Wu, Hang Ye, and Yuntian Gu. Guided diffusion model for adversarial purification from random noise. *arXiv preprint arXiv:2206.10875*, 2022.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. Shadowcast: Stealthy data poisoning attacks against vision-language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024a.

Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. Cross-modality information check for detecting jailbreaking in multimodal large language models. In *Findings of the Association for Computational Linguistics*, pp. 13715–13726, 2024b.

Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Computer Vision and Pattern Recognition*, pp. 2922–2931, 2023a.

Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pre-training against data poisoning and backdoor attacks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 10678–10691, 2023b.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, Yuxin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms. *arXiv preprint arXiv:2405.13800*, 2024.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *In Proceedings of the 10th International Conference on Learning Representations ICLR,*, 2022.

Nir Yellinek, Leonid Karlinsky, and Raja Giryes. 3VL: using trees to teach vision & language models compositional concepts. *arXiv preprint arXiv:2312.17345*, 2023.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. VLATTACK: Multimodal adversarial attacks on vision-language tasks via pre-trained models. In *Conference on Neural Information Processing Systems*, 2023.

Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pp. 12062–12072. PMLR, 2021.

Chung-En Johnny Yu, Brian Jalaian, Nathaniel D Bastian, et al. Orca: Agentic reasoning for hallucination and adversarial robustness in vision-language models. *arXiv preprint arXiv:2509.15435*, 2025.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 1307–1315, 2018.

Lu Yu, Haiyang Zhang, and Changsheng Xu. Text-guided attention is all you need for zero-shot robustness in vision-language models. In *In Proceedings of the Advances in neural information processing systems*, 2024.

Zhen Yu, Zhou Qin, Zhenhua Chen, Meihui Lian, Haojun Fu, Weigao Wen, Hui Xue, and Kun He. Sparse black-box multimodal attack for vision-language adversary generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5775–5784, 2023.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistic*, 2024a.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the ACM International Conference on Multimedia*, pp. 5005–5013, 2022.

Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *Proceedings of the European Conference on Computer Vision*, 2023.

Peng-Fei Zhang, Zi Huang, and Guangdong Bai. Universal adversarial perturbations for vision-language pre-trained models. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 862–871, 2024b.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 5579–5588, 2021a.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *Proceedings of ICLR*, 2024c.

Xingwei Zhang, Xiaolong Zheng, and Wenji Mao. Adversarial perturbation defense on deep neural networks. *ACM Computing Surveys (CSUR)*, 54(8):1–36, 2021b.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. Benchmarking trustworthiness of multimodal large language models: A comprehensive study. *arXiv preprint arXiv:2406.07057*, 2024d.

Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, and Yu Wang. Pip: Detecting adversarial examples in large vision-language models via attention patterns of irrelevant probe questions. In *ACM International Conference on Multimedia*, pp. 11175–11183, 2024e.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *European Conference on Computer Vision*, pp. 127–142. Springer, 2025.

Tianyi Zhao, Liangliang Zhang, Yao Ma, and Lu Cheng. A survey on safe multi-modal learning systems. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6655–6665, 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 54111–54138, 2023.

Haonan Zheng, Xinyang Deng, Wen Jiang, and Wenrui Li. A unified understanding of adversarial vulnerability regarding unimodal models and vision-language pre-training models. In *ACM Multimedia*, 2024a.

Haonan Zheng, Wen Jiang, Xinyang Deng, and Wenrui Li. Sample-agnostic adversarial perturbation for vision-language pre-training models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9749–9758, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena, 2023.

Jinghao Zhou, Li Dong, Zhe Gan, Lijuan Wang, and Furu Wei. Non-contrastive learning meets language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11028–11038, 2023a.

Wanqi Zhou, Shuanghao Bai, Qibin Zhao, and Badong Chen. Revisiting the adversarial robustness of vision language models: a multimodal perspective. *arXiv preprint arXiv:2404.19287*, 2024a.

Yiwei Zhou, Xiaobo Xia, Zhiwei Lin, Bo Han, and Tongliang Liu. Few-shot adversarial prompt learning on vision-language models. In *Proceedings of the Neural Information Processing Systems*, 2024b.

Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. AdvCLIP: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *ACM International Conference on Multimedia*, pp. 6311–6320, 2023b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Xiaohua Zhu, Qibin Zhao, Weijie Xie, et al. Vlmo: Unified vision-language pre-training with masked vision-language modeling. In *International Conference on Learning Representations*, 2021.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *The 41st International Conference on Machine Learning*, 2024a.

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. In *International Conference on Machine Learning*, 2024b.