DISCRETE TEXT-CONDITIONED UNIVERSAL MULTIMODAL ATTACKS ON VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language pre-trained (VLP) models support diverse multimodal tasks but remain vulnerable to universal adversarial perturbations. Continuous textual surrogates can create a mismatch between optimization and the discrete edits applied at inference, while a shared word has context-dependent effects. We propose an optimization-based framework that combines online vocabulary-constrained text learning with edited-caption conditioning: the current shared word is inserted into complete captions, which are re-encoded to condition universal image-perturbation learning. Representation-level disruption and bidirectional cross-modal competition target image–text correspondence, with auxiliary fusion guidance when supported by the source architecture. Experiments span six VLP configurations and three vision–language tasks. Extensive experiments demonstrate competitive white-box attack performance, and effective transfer to downstream vision–language tasks without further optimization.

1 INTRODUCTION

Vision–language pre-trained (VLP) models support increasingly diverse applications (Radford et al., 2021; Singh et al., 2022; Li et al., 2023), yet remain vulnerable to constrained adversarial modifications (Zhang et al., 2022; Lu et al., 2023; Yin et al., 2023). Universal adversarial perturbations (UAPs) expose vulnerabilities shared across inputs by reusing attack content rather than optimizing a separate attack for every sample (Moosavi-Dezfooli et al., 2017; Wallace et al., 2019). In multimodal attacks, a bounded image perturbation is combined with a reusable word (Fang et al., 2025; Ye et al., 2026). The word is shared, but its effect depends on the sentence in which it appears. This distinction makes it important to separate the continuous representation used for textual optimization, the discrete edit applied to the input, and the text features used to learn the image perturbation.

Recent work has progressed from visual UAPs to attacks on both modalities. ETU strengthens image perturbations through diversified visual contexts (Zhang et al., 2024). C-PGC uses cross-modal-conditioned generators and converts generated textual embeddings into vocabulary items (Fang et al., 2025). HRA refines image-perturbation learning and ranks influential corpus words (Zhang & Huang, 2026). DO-UAP directly optimizes multimodal perturbations using perturbed images and text (Yang et al., 2024), while ARG-Attack introduces asymmetric geometric and relational objectives (Ye et al., 2026). Thus, both joint optimization and relational disruption have established precedents.

Our focus is the connection between discrete textual realization and image-side learning. C-PGC's image-generator algorithm uses original caption embeddings, and HRA separates image optimization from trigger selection (Fang et al., 2025; Zhang & Huang, 2026). We combine online vocabulary projection with *edited-caption conditioning*: inserting the current shared word into each complete caption and re-encoding that caption to obtain the condition for image updates. The attack word remains shared across samples. Under these conditions, bidirectional competition targets strong remaining correct associations, and fusion-based source models provide additional internal-response objectives. These auxiliary signals complement representation displacement. Our contributions are as follows:

- We connect online vocabulary-constrained text learning with edited-caption-conditioned image optimization, learning one shared image perturbation and one reusable attack word.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- We introduce bidirectional competition and architecture-aware fusion guidance as auxiliary objectives targeting relative image–text associations and accessible internal responses.
- We evaluate the proposed framework across six VLP model configurations, multiple datasets, and downstream vision-language tasks, assessing attack effectiveness and the transfer of fixed perturbations across models and tasks without task-specific re-optimization.

2 RELATED WORK

2.1 VISION-LANGUAGE PRE-TRAINING MODELS

VLP research spans cross-modal encoders (Lu et al., 2019; Tan & Bansal, 2019; Chen et al., 2020; Li et al., 2020b; Kim et al., 2021), contrastive alignment (Jia et al., 2021; Yao et al., 2022), and unified multimodal pre-training (Wang et al., 2023a). We study five families providing unimodal image and text features: CLIP uses independent encoders with contrastive alignment (Radford et al., 2021), whereas ALBEF, TCL, X-VLM, and BLIP additionally contain cross-modal fusion branches (Li et al., 2021; Yang et al., 2022; Zeng et al., 2022; Li et al., 2022). This distinction determines which attack signals are accessible from the source model.

2.2 ADVERSARIAL ATTACKS ON VISION-LANGUAGE MODELS

Adversarial attacks evaluate model robustness by introducing constrained modifications to inputs and measuring the resulting prediction failures (Goodfellow et al., 2015; Moosavi-Dezfooli et al., 2016; Madry et al., 2018).Transfer-oriented visual attacks improve cross-model generalization through momentum accumulation, input diversity, and translation-invariant optimization (Dong et al., 2018; Xie et al., 2019; Dong et al., 2019).Textual attacks construct discrete edits through gradient-guided search (Ebrahimi et al., 2018) or constrained word substitutions (Jin et al., 2020); contextual language-model predictions further guide adversarial text construction (Li et al., 2020a; Garg & Ramakrishnan, 2020).Unlike unimodal classification attacks, attacks on VLP models must account for both modality-specific representations and the correspondence between images and text. Co-Attack extended white-box attacks to image-text inputs and explicitly incorporated cross-modal relationships into perturbation construction (Zhang et al., 2022). Subsequent work placed greater emphasis on transferability. SGA used augmented image-text sets and set-level cross-modal guidance to improve transfer to inaccessible models (Lu et al., 2023), while TMM exploited modality-consistency and modality-discrepancy information to construct transferable multimodal examples (Wang et al., 2024b). Other VLP attacks improve transferability through self-augmentation and trajectory diversification (He et al., 2023; Gao et al., 2024), as well as joint contrastive optimization and collaborative cross-modal interaction (Wang et al., 2023b; Fu et al., 2024).These multimodal attacks remain instance-specific, requiring a new adversarial example for each input pair despite their cross-model transferability.

2.3 UNIVERSAL MULTIMODAL ADVERSARIAL PERTURBATIONS

Earlier visual UAP methods explore data-free feature disruption and hidden-layer sensitivity (Mopuri et al., 2017; 2019; Khrulkov & Oseledets, 2018), generative synthesis (Mopuri et al., 2018), and improved generalization through gradient aggregation and feature disruption (Liu et al., 2023; Wang et al., 2024a).Visual-only universal attacks include AdvCLIP, which learns downstream-agnostic adversarial patches, and ETU, which strengthens image perturbations through global and local objectives and diversified visual contexts (Zhou et al., 2023; Zhang et al., 2024). Multimodal methods additionally construct reusable textual modifications. C-PGC trains cross-modal-conditioned generators and maps generated textual embeddings to vocabulary items (Fang et al., 2025). HRA selects corpus words through intra- and inter-sentence importance while refining image-perturbation learning (Zhang & Huang, 2026). DO-UAP directly optimizes multimodal perturbations using cross-modal and unimodal objectives with selected image augmentations (Yang et al., 2024). ARG-Attack combines geometric objectives with relational polarity to disrupt matched and mismatched associations (Ye et al., 2026).

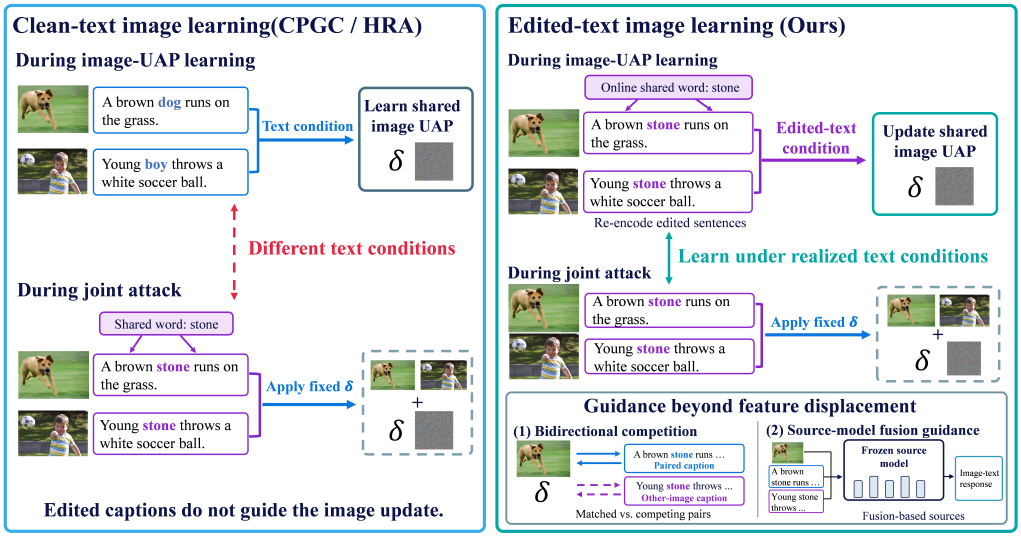

Figure 1: Text conditions in universal multimodal attacks. Left: the image-learning paths of C-PGC and HRA use original captions. Right: our image updates use re-encoded edited captions. The inset shows bidirectional competition and source-dependent fusion guidance.

These methods differ in textual realization, image-side conditions, and relational objectives. C-PGC's image-generator algorithm uses original caption embeddings, whereas HRA separates image optimization from trigger selection. DO-UAP already processes perturbed images and text jointly. Our framework combines online source-vocabulary projection with edited-caption conditioning of image updates. ARG-Attack's relational polarity loss contrasts average matched and mismatched similarities, whereas our bidirectional competition targets the strongest correct associations against the hardest mismatches.

3 Method

3.1 Problem and Framework

Given paired images and captions $\mathcal{D}_s = \{(x_i, \mathcal{T}_i)\}_{i=1}^{N}$ and a frozen source VLP model with encoders f_v, f_t, we learn a shared image perturbation δ and a reusable word $w^\star$:

$$\widehat{x} = \mathrm{clip}_{[0,1]}(x + \delta), \quad \|\delta\|_\infty \le \epsilon_v, \qquad \widehat{t} = \mathcal{R}(t, w^\star, p(t)), \quad d_{\mathrm{sub}}(t, \widehat{t}) \le \epsilon_t = 1. \tag{1}$$

Here, $\mathcal{R}$ replaces the word at position $p(t)$, and d_{sub} counts replacements, following C-PGC's budget (Fang et al., 2025). Only the position varies across inputs; $(\delta, w^\star)$ is shared. This budget does not itself guarantee semantic preservation.

Figure 2 links two alternating updates. A continuous parameter z learns the shared word through online vocabulary projection. The updated word is then inserted into complete captions, which are re-tokenized and encoded to condition image learning. This *edited-caption conditioning* connects discrete edits with image optimization without learning sample-specific words.

3.2 Online Discrete Text Learning

Mapping a continuous textual surrogate to a word only after learning can create an optimization–execution mismatch (Fang et al., 2025). We instead search a source-vocabulary set $E_s = \{(e_k, w_k)\}_{k=1}^{K}$, with embeddings $e_k \in \mathbb{R}^{d_e}$, retaining valid single-token words whose decoding and re-encoding recover the same token. Let $\bar{z}$ be z with its norm clipped to the candidate norm range. Online projection uses

$$k^\star = \arg\min_k \|\bar{z} - e_k\|_2^2, \qquad w^\star = w_{k^\star}, \qquad \widetilde{e} = \bar{z} + \mathrm{sg}(e_{k^\star} - \bar{z}), \tag{2}$$

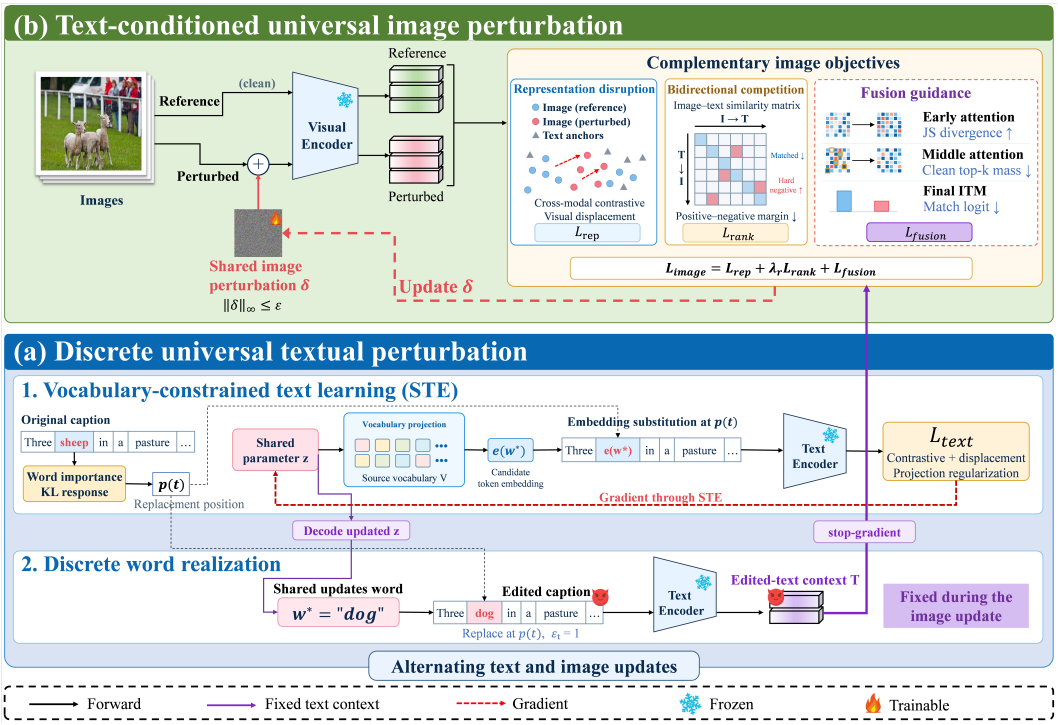

Figure 2: *Overview of the proposed framework.* **(a)** A shared textual parameter is optimized through vocabulary projection and a straight-through estimator. The updated shared word replaces the original word at a context-dependent position, providing fixed text context for the subsequent image update. **(b)** A shared image perturbation is jointly optimized by representation disruption and bidirectional competition. For source models with an explicit fusion pathway, additional guidance is provided through changes in attention distribution. All source-model parameters remain frozen, while gradients update the shared perturbation parameters.

where sg stops gradients. The forward value is a candidate embedding; the backward pass uses a straight-through approximation (Bengio et al., 2013). Following C-PGC, $p(t)$ selects the valid word whose masking maximizes KL divergence between the softmax-normalized original and masked sentence features.

All global features below are ℓ_2-normalized. For B images and M captions, $\pi(j)$ identifies caption j's paired image. Let $u_j(z)$ be its feature under STE embedding replacement, t_j^0, c_i^0 the original caption/image features, and $\bar{c}_r$ unmodified image features from a separately shuffled reference batch. With $r_j^\star = \arg\min_r u_j^\top \bar{c}_r$ and temperature τ, define

$$A_t = \frac{1}{M}\sum_j e^{u_j^\top c^0_{\pi(j)}/\tau}, \qquad B_t = \frac{1}{M}\sum_j e^{u_j^\top \bar{c}_{r_j^\star}/\tau},$$
$$\mathcal{L}_{\text{text}} = \log\frac{A_t}{A_t + B_t} - \alpha\sum_j \|u_j - t_j^0\|_2^2 + \lambda_c d_1 + \lambda_p[d_1 - d_2 + m_p]_+. \tag{3}$$

Here d_1, d_2 are the two smallest squared candidate distances $\|\bar{z} - e_k\|_2^2/d_e$, and $[a]_+ = \max(a, 0)$. Minimizing $\mathcal{L}_{\text{text}}$ over z combines association disruption, summed textual displacement, and projection regularization; clean/reference features stay fixed. STE embedding replacement is used only for this update, while image learning uses re-encoded edited captions.

3.3 Edited-Caption-Conditioned Image Learning

A shared word has sentence-dependent effects. We therefore fix $t_j^a = f_t(\widehat{t_j})$, re-encoded after each text update, as the image-side condition. For five augmented views, $c_i^{(v)}$ and $a_i^{(v)}$ denote features

before and after adding the same δ. Image learning minimizes

$$\min_{\|\delta\|_\infty \le \epsilon_v} \mathcal{L}_{\text{image}}, \qquad \mathcal{L}_{\text{image}} = \underbrace{\mathcal{L}_{\text{rep}} + \lambda_r \mathcal{L}_{\text{rank}}}_{\mathcal{L}_{\text{base}}} + \mathcal{L}_{\text{fusion}}. \tag{4}$$

Representation disruption. We retain C-PGC's matched/distant-reference construction (Fang et al., 2025), replacing the matched text conditions with edited captions. Let $\mathcal{P}(i) = \{j : \pi(j) = i\}$ and $\mathcal{Q}_r$ be image r's captions in a separately shuffled reference batch, with unmodified features $\bar{t}_l$. The clean view selects $r_i^{(v)} = \arg\min_r |\mathcal{Q}_r|^{-1} \sum_{l \in \mathcal{Q}_r} (c_i^{(v)})^\top \bar{t}_l$. Then

$$\begin{aligned} M_v &= \frac{1}{B} \sum_i \sum_{j \in \mathcal{P}(i)} e^{(a_i^{(v)})^\top t_j^a / \tau}, \qquad R_v = \frac{1}{B} \sum_i \sum_{l \in \mathcal{Q}_{r_i^{(v)}}} e^{(a_i^{(v)})^\top \bar{t}_l / \tau}, \\ \mathcal{L}_{\text{rep}} &= \frac{1}{5} \sum_{v=1}^{5} \left[\log \frac{M_v}{M_v + R_v} + \lambda_v C_{\text{vis}}^{(v)} \right]. \end{aligned} \tag{5}$$

$C_{\text{vis}}^{(v)}$ averages clean/adversarial cosine agreement over samples and local positions: fusion sources and CLIP RN101 equally weight global/local readouts, while CLIP ViT-B/16 weights global, CLS, and patch readouts equally. Minimization encourages visual displacement; clean, reference, and edited-caption features remain fixed.

Bidirectional competition. To weaken relative correspondence beyond aggregate similarities, we compare the strongest correct association with the hardest mismatch. For $S_{ij}^{(v)} = (a_i^{(v)})^\top t_j^a$, define $\phi(g) = \tau_r \log(1 + e^{(g+m_r)/\tau_r})$ and

$$\begin{aligned} g_i^{v,\text{I2T}} &= \max_{j \in \mathcal{P}(i)} S_{ij}^{(v)} - \max_{j \notin \mathcal{P}(i)} S_{ij}^{(v)}, \qquad g_j^{v,\text{T2I}} = S_{\pi(j)j}^{(v)} - \max_{i \neq \pi(j)} S_{ij}^{(v)}, \\ \mathcal{L}_{\text{rank}} &= \frac{1}{10} \sum_{v=1}^{5} \left[\frac{1}{B} \sum_i \phi(g_i^{v,\text{I2T}}) + \frac{1}{M} \sum_j \phi(g_j^{v,\text{T2I}}) \right]. \end{aligned} \tag{6}$$

Minimization encourages mismatches to overtake correct associations in both directions.

3.4 Fusion Guidance

Fusion-based sources additionally expose cross-attention and image–text matching (ITM) responses. At step s (zero-based), we use view $v = 1 + (s \bmod 5)$ with the same edited captions in its clean/adversarial paths. Attention P_{jlhqu}^b indexes caption, layer, head, text query, and visual key, with $b \in \{0, \delta\}$. Values are clamped below at 10^{-8} and normalized over all visual keys, including CLS. Valid queries exclude padding/boundary tokens, falling back to non-padding tokens if empty. Split the ordered layers into three nearly equal consecutive groups; $\mathcal{G}_E, \mathcal{G}_M$ denote the first two, and $\langle \cdot \rangle_{\mathcal{G}}$ averages heads, valid queries, and layers.

Let $\mathcal{K}_{jlhq}$ hold the clean attention's top-$\min(5, U)$ keys, where U is the number of visual keys, and $b_j^b = \langle \sum_{u \in \mathcal{K}_{jlhq}} P_{jlhqu}^b \rangle_{\mathcal{G}_M}$. With matched-class ITM logits c_j^b, the three responses are

$$\begin{aligned} D_{E,j} &= \text{clip}_{[0,1]} \left(\langle \text{JS}(P_{jlhq}^0 \| P_{jlhq}^\delta) \rangle_{\mathcal{G}_E} / \log 2 \right), \\ D_{M,j} &= \text{clip}_{[-1,1]} \left(\frac{b_j^0 - b_j^\delta}{\max(|b_j^0|, 10^{-6})} \right), \\ D_{\text{itm},j} &= \tanh \left(\frac{c_j^0 - c_j^\delta}{\max(\text{Std}_{\text{pop}}(\{c_j^0\}_{j=1}^M), 1)} \right). \end{aligned} \tag{7}$$

Clean responses are detached; the ITM scale is the population standard deviation across current pairs.

Table 1: ASR (%) of different methods for image–text retrieval on Flickr30k. TR: text retrieval given the input image; IR: image retrieval given the input text. $*$: the model variant without the FUSION GUIDANCE module.

Source	Method	ALBEF		TCL		X-VLM		$\text{CLIP}_{\text{ViT-B/16}}$		$\text{CLIP}_{\text{RN101}}$		BLIP	
		TR	IR	TR	IR	TR	IR	TR	IR	TR	IR	TR	IR
ALBEF	GAP	69.78	81.59	22.15	29.97	6.61	18.37	23.40	37.54	29.92	44.29	16.09	28.12
	ETU	78.01	84.56	29.92	35.91	14.33	22.03	23.77	39.20	33.55	47.69	22.61	32.28
	CPGC	90.13	88.82	62.11	64.48	20.53	39.38	43.10	65.93	**54.40**	72.51	44.79	56.36
	ARG-Attack	83.56	85.41	62.84	69.25	27.44	50.86	42.98	68.33	**54.40**	75.77	55.63	66.63
	Ours	**96.40**	**96.41**	**70.60**	**80.60**	**28.25**	**53.42**	**43.23**	**72.44**	50.13	**78.57**	**60.67**	**71.21**
TCL	GAP	33.50	40.61	82.41	80.67	6.61	17.79	21.55	38.56	30.57	45.48	21.45	31.82
	ETU	28.26	35.03	90.48	87.57	9.65	20.56	25.00	39.68	36.14	49.33	18.93	29.19
	CPGC	50.26	56.29	94.93	90.64	14.94	33.96	46.92	66.41	52.98	70.66	35.75	52.52
	ARG-Attack	53.44	66.60	88.92	86.67	28.35	**51.51**	**52.71**	73.05	**59.84**	77.77	**53.52**	69.03
	Ours	**54.98**	**71.46**	**95.24**	**94.29**	**28.76**	50.58	51.23	**79.17**	56.09	**80.32**	52.47	**73.91**
X-VLM	GAP	16.14	24.43	17.08	26.20	90.24	85.98	24.51	41.15	42.62	53.08	16.19	25.74
	ETU	12.33	21.93	13.98	24.04	93.19	90.85	23.89	39.62	35.62	51.19	12.09	23.59
	CPGC	24.46	47.77	29.19	50.15	93.29	91.90	43.47	66.03	**59.20**	72.79	32.39	52.24
	ARG-Attack	**29.29**	49.22	**36.54**	52.73	**96.54**	91.97	44.58	66.54	57.90	72.83	**35.86**	53.63
	Ours	25.18	**52.16**	28.47	**55.96**	95.33	**94.23**	**49.38**	**70.62**	56.74	**75.60**	33.44	**60.13**
$\text{CLIP}_{\text{ViT-B/16}}$	GAP	11.72	23.34	15.32	26.39	8.54	20.48	85.73	90.45	48.83	60.78	14.83	26.46
	ETU	14.80	25.23	21.22	30.87	10.87	24.96	84.14	90.45	57.51	65.51	16.40	27.22
	CPGC	23.23	38.67	25.05	41.79	15.85	35.59	88.92	**93.05**	66.06	75.42	26.71	45.70
	ARG-Attack	**24.25**	40.96	**30.75**	44.48	20.53	38.98	89.53	92.34	**70.21**	79.34	**28.81**	**48.29**
	Ours*	21.58	**43.40**	27.43	**47.2**	**20.63**	**43.32**	**96.43**	91.16	65.93	**79.87**	26.81	47.27
$\text{CLIP}_{\text{RN101}}$	GAP	13.57	25.21	19.05	28.87	11.59	23.13	27.46	43.16	73.18	81.60	15.25	27.94
	ETU	8.94	20.59	13.25	24.41	8.94	20.82	21.92	40.51	91.71	92.40	11.15	23.82
	CPGC	19.01	41.86	22.98	47.02	19.61	43.26	40.89	65.77	**96.50**	94.22	24.19	48.17
	ARG-Attack	**26.52**	41.23	25.78	42.83	25.41	42.31	50.25	69.92	90.67	92.58	**29.23**	**51.11**
	Ours*	24.87	**44.87**	**30.64**	**50.03**	**25.71**	**46.58**	**56.28**	68.29	95.47	**94.71**	26.39	49.81
BLIP	GAP	12.23	23.94	14.49	25.44	6.91	17.81	20.32	37.00	26.81	43.59	47.21	73.33
	ETU	19.32	27.91	19.98	29.15	11.99	20.91	24.38	39.84	31.61	46.22	59.52	77.82
	CPGC	32.17	44.40	33.44	44.51	**18.60**	**35.53**	43.35	60.26	48.96	66.95	71.82	82.82
	ARG-Attack	34.53	47.95	38.41	49.34	15.65	32.52	**46.43**	**62.68**	**54.53**	**70.10**	62.15	73.24
	Ours	**39.16**	**56.74**	**42.24**	**51.52**	15.75	33.53	38.79	58.98	48.06	64.22	**86.23**	**87.3**

4 EXPERIMENTS

We evaluate universal perturbations along three axes: attack effectiveness across VLP architectures, transfer to downstream tasks, and the contribution of each component.

4.1 EXPERIMENTAL SETUP

Datasets and protocols. We evaluate image–text retrieval on Flickr30K (Young et al., 2014), image captioning on MSCOCO (Lin et al., 2014), and visual grounding on RefCOCO+ (Yu et al., 2016). Retrieval uses the official C-PGC annotations (Fang et al., 2025): 30,000 images and 150,000 captions for perturbation learning, and 1,000 images and 5,000 captions for evaluation.

Models and baselines. We compare against GAP, a generator-based visual attack (Poursaeed et al., 2018); ETU, a visual UAP method for VLP models (Zhang et al., 2024); C-PGC, a generative multimodal attack (Fang et al., 2025); and ARG-Attack, a directly optimized multimodal attack (Ye et al., 2026). The retrieval suite comprises six configurations from five VLP families: ALBEF (Li et al., 2021), TCL (Yang et al., 2022), X-VLM (Zeng et al., 2022), BLIP (Li et al., 2022), and CLIP (Radford et al., 2021) with ViT-B/16 (Dosovitskiy et al., 2021) and RN101 (He et al., 2016) backbones. The four fusion-based families support the auxiliary fusion objective; CLIP does not. Downstream targets are BLIP-base for captioning and ALBEF and X-VLM for grounding.

Attack constraints and metrics. We learn a separate shared image–text perturbation pair for each frozen source model using only its gradients. Following C-PGC, we constrain the image perturbation by $\|\delta\|_\infty \leq \epsilon_v = 12/255$ in $[0, 1]$ pixel space and allow at most one word substitution per sentence

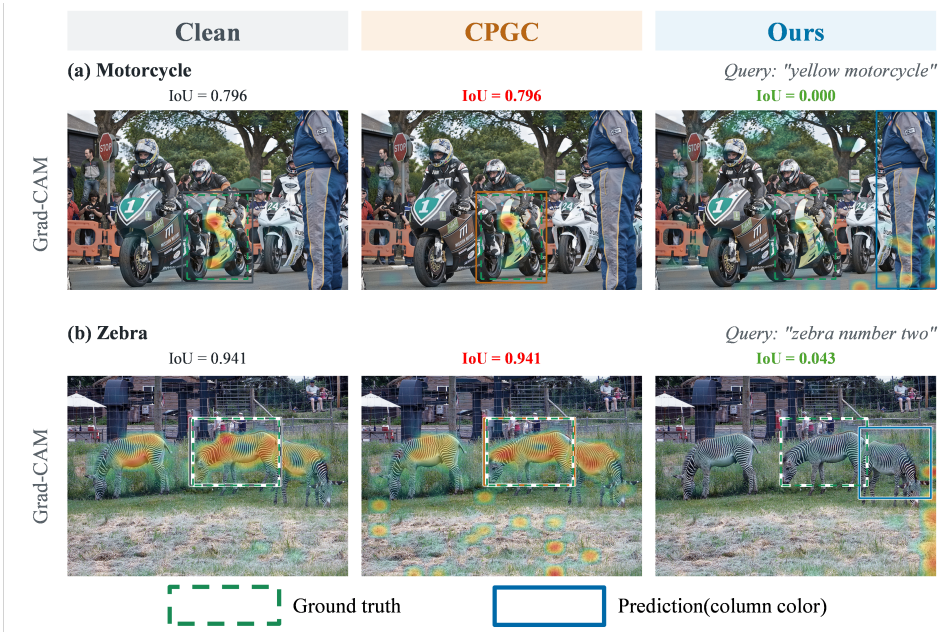

Figure 3: Visual grounding on RefCOCO+ with ALBEF as source and target. The middle column (*Other Uap*) is C-PGC. Ground-truth and predicted boxes are overlaid on the original images; queries and IoUs correspond to the evaluated clean or joint-attack inputs.

($\epsilon_t = 1$). The replacement word is shared across inputs, with a sentence-dependent replacement position. For retrieval, we report attack success rate (ASR) in both image-to-text (I2T) and text-to-image (T2I) directions, using $K = 1$ for the main comparison and additionally $K \in \{5, 10\}$ for ablations:

$$\mathrm{ASR}_d@K = 100 \frac{\sum_q \mathbf{1}[r_d^{\mathrm{clean}}(q) \le K]\, \mathbf{1}[r_d^{\mathrm{adv}}(q) > K]}{\sum_q \mathbf{1}[r_d^{\mathrm{clean}}(q) \le K]}. \tag{8}$$

Here, $r_d(q)$ is the one-based rank of the highest-ranked correct match for query q in direction d; for I2T, this is the minimum rank across all paired captions. ASR measures the fraction of initially successful queries that fail under attack, with higher values indicating stronger attacks.

Implementation details. We train for 20 epochs with seed 42 and a global batch size of 24 on four NVIDIA H20 GPUs. Losses are computed within each device's mini-batch, and gradients are averaged across devices. Image and text variables use separate Adam optimizers (Kingma & Ba, 2015) with cosine learning-rate decay, from $0.5/255$ to $0.05/255$ and from 0.005 to 0.0005, respectively. Following C-PGC, the five image views use scales $\{0.5, 0.75, 1, 1.25, 1.5\}$: the base view includes uniform noise in $[-\epsilon_v, \epsilon_v]$, and the four additional views receive Gaussian noise $\mathcal{N}(0, 0.05^2)$ before rescaling.

4.2 Universal Attack Effectiveness

On Flickr30K, each source-model perturbation is evaluated against all six retrieval configurations; black-box means exclude the source. Table 1 shows that our framework attains the highest white-box ASR@1 in nine of twelve source–direction settings. Across the 30 black-box source–target pairs, mean I2T/T2I ASR@1 is 40.00%/60.39%, compared with 35.82%/53.41% for C-PGC and 40.55%/57.53% for ARG-Attack. The improvement is therefore clearer for T2I transfer, while mean I2T remains slightly below ARG-Attack. Performance varies with the model pair: X-VLM-to-TCL I2T ASR@1 is 28.47%, versus 36.54% for ARG-Attack, and BLIP-to-CLIP RN101 transfer also trails ARG-Attack in both directions. These results distinguish strong source-model attacks from source- and direction-dependent transfer.

4.3 TRANSFER TO DOWNSTREAM TASKS

We apply the Flickr30K-learned perturbations unchanged to downstream checkpoints, without task-specific optimization.

Visual grounding. We apply the learned image–text perturbations to ALBEF and X-VLM grounding targets on RefCOCO+, reporting accuracy at IoU ≥ 0.5.ALBEF uses Grad-CAM-based localization (Selvaraju et al., 2017), whereas X-VLM uses its native localization output. As shown in Table 2, our method yields lower attacked accuracy than C-PGC in 16 of 18 comparisons, demonstrating effective cross-task transfer.Figure 3 illustrates the resulting localization errors in two examples.

Table 2: Transfer attack performance. Image–text perturbations trained on Flickr30k with different source models are transferred to attack a fixed ALBEF grounding target. Accuracy is reported at IoU ≥ 0.5 on RefCOCO+ Val/TestA/TestB.

Target	Method	Baseline			ALBEF			TCL			X-VLM		
		Val	TestA	TestB	Val	TestA	TestB	Val	TestA	TestB	Val	TestA	TestB
ALBEF	CPGC	58.44	65.93	46.25	32.15	33.85	29.84	43.65	47.90	37.41	41.16	48.97	**31.36**
	Ours	58.44	65.93	46.25	**28.71**	**30.14**	**26.94**	**39.12**	**42.25**	**34.06**	**37.65**	**41.53**	32.60
X-VLM	CPGC	84.51	89.03	76.93	59.59	61.40	56.58	62.13	63.11	59.42	10.83	14.20	5.95
	Ours	84.51	89.03	76.93	**50.68**	**48.94**	**50.73**	**55.88**	**54.98**	**56.86**	**7.26**	**7.06**	**7.63**

Table 3: Cross-task transfer of Flickr30K-learned image perturbations to BLIP-base image captioning on MSCOCO. Clean denotes unperturbed performance. Lower scores indicate stronger attacks; bold marks the lowest score in each metric column.

Source	Method	B@4 ↓	METEOR ↓	ROUGE-L ↓	CIDEr ↓	SPICE ↓
Baseline	–	39.65	30.98	59.97	133.31	23.81
ALBEF	CPGC	28.52	24.55	51.09	92.63	17.54
	Ours	25.11	22.99	48.16	85.83	16.17
TCL	CPGC	28.42	24.93	51.03	97.06	18.02
	Ours	29.32	25.11	51.34	96.81	18.02
BLIP	CPGC	26.62	23.91	49.77	85.50	16.94
	Ours	**14.82**	**17.65**	**39.38**	**49.54**	**10.72**



Clean: a kitchen with a sink and a window
C-PGC: a reflection of a window with curtains in it
Ours: a group of people sitting around a living room



Clean: a wooden box filled with lots of ripe and unripe bananas
C-PGC: a poster with a bunch of bananas on it
Ours: a man riding a bike down a dirt road

Figure 4: Image-captioning examples on MSCOCO with BLIP-base as target. Captions are generated from clean inputs or inputs with Flickr30K-learned image perturbations. Original images are shown for reference.

Image captioning. We evaluate BLIP-base on the 5,000-image MSCOCO Karpathy test split (Karpathy & Fei-Fei, 2015), applying only the image perturbation.We report BLEU-4 (Papineni et al., 2002),METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam

et al., 2015), and SPICE (Anderson et al., 2016); lower scores indicate stronger attacks.As shown in Table 3, the BLIP-source perturbation produces the largest degradation across all five metrics, reducing CIDEr from 133.31 to 49.54, versus 85.50 for C-PGC.ALBEF-source perturbations also outperform C-PGC on all five metrics,while TCL-source results are comparable, indicating source-dependent cross-task transfer.Figure 4 shows examples of generated captions under the transferred perturbations.

4.4 Ablation Studies

Using ALBEF as the frozen source on Flickr30K, we ablate *discrete text coupling* edited-caption conditioning(D), *bidirectional competition* (R), and *fusion guidance* (F). A0 uses continuous text and representation objectives; A1–A4 add the components in Table 4. All variants use the same schedule and budgets and are evaluated at the final checkpoint. Black-box means equally weight the five non-source targets in Table 5.

Table 4: Component ablations on Flickr30K with ALBEF as the source. D, R, and F are defined in the text. Entries are I2T/T2I ASR (%, ↑); black-box means equally weight the five non-source targets. Bold marks the best value per direction and column. Means and differences are computed before final rounding.

				ALBEF			Black-box mean		
Variant	D	R	F	@1	@5	@10	@1	@5	@10
A0	–	–	–	83.56/90.79	69.14/84.61	62.60/81.95	38.33/56.98	17.68/34.57	11.18/26.59
A1	✓	–	–	94.35/95.40	88.88/91.47	86.50/89.24	50.95/**71.32**	28.88/51.21	21.78/42.63
A2	✓	✓	–	95.79/96.10	91.78/92.87	90.10/90.64	50.60/71.06	27.99/50.63	20.62/41.91
A3	✓	–	✓	94.96/95.45	90.98/91.65	88.60/89.62	**51.48**/71.31	**29.84/51.24**	**22.46/42.80**
A4	✓	✓	✓	**96.40/96.41**	**92.79/93.74**	**91.10/91.63**	50.58/71.25	28.69/51.21	21.42/42.19

Component effects. Adding D yields the largest improvement over A0: A1 raises mean black-box I2T/T2I ASR@1 from 38.33/56.98 to 50.95/71.32, with gains also observed at $K = 5$ and $K = 10$. The auxiliary objectives have different effects. Adding R (A2) raises source-model ASR@1 by 1.44/0.70 percentage points over A1; adding F (A3) gives the highest mean black-box I2T ASR@1 and the strongest mean transfer at $K = 5$ and $K = 10$ in both directions.

Table 5: Target-wise black-box ASR@1 (%, ↑), reported as I2T/T2I. CLIP-ViT denotes ViT-B/16. Bold marks the best value per direction and target.

Variant	TCL	BLIP	X-VLM	CLIP-ViT	CLIP-RN101
A0	49.07/57.38	35.44/51.57	18.50/36.02	42.12/65.20	46.50/74.72
A1	70.60/79.79	59.31/**71.60**	27.74/53.17	**46.06**/72.09	51.04/**79.97**
A2	71.22/80.56	57.83/69.66	**29.27/53.75**	42.61/71.90	**52.07**/79.45
A3	**73.81**/80.56	**60.99**/71.37	27.95/53.04	43.47/71.96	51.17/79.62
A4	70.60/**80.60**	60.67/71.21	28.25/53.42	43.23/**72.44**	50.13/78.57

Combined framework. A4 gives the highest source-model ASR at all three thresholds, reaching 96.40/96.41 at $K = 1$, while its mean black-box ASR@1 remains below A1 and A3. The ablations therefore distinguish stronger source-model attacks from stronger average transfer

5 Conclusion

We presented a universal multimodal attack framework combining online discrete text learning, edited-caption-conditioned image optimization, and auxiliary competition and fusion objectives. Experiments support its attack effectiveness and cross-task transfer in the evaluated settings, while showing that stronger source-model attacks do not necessarily yield stronger average transfer.

AI USE STATEMENT

We use ChatGPT to assist in drafting and polishing the manuscripts, translation, and literature organization. The author has reviewed and revised the manuscripts completed with the aid of AI, and is responsible for verifying their consistency with the implementation process and the cited literature. The author is fully responsible for the final content, including all methods, results, and conclusions.

ETHICS STATEMENT

This work investigates universal adversarial perturbations to assess the robustness of vision–language models using established benchmark datasets and pretrained checkpoints. The methods have dual-use potential, as reusable perturbations could be misused to disrupt multimodal systems. Our research aims to support authorized robustness evaluation and future defensive research. We advocate responsible use of the methods and experimental findings.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, volume 9909 of *Lecture Notes in Computer Science*, pp. 382–398. Springer, 2016. doi: 10.1007/978-3-319-46454-1_24. URL https://link.springer.com/chapter/10.1007/978-3-319-46454-1_24.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. URL https://arxiv.org/abs/1308.3432. arXiv:1308.3432.

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 *Computer Vision – ECCV 2020*, volume 12375 of *Lecture Notes in Computer Science*, pp. 104–120. Springer, 2020. doi: 10.1007/978-3-030-58577-8_7. URL https://link.springer.com/chapter/10.1007/978-3-030-58577-8_7.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Dong_Evading_Defenses_to_Transferable_Adversarial_Examples_by_Translation-Invariant_Attacks_CVPR_2019_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-2006. URL https://aclanthology.org/P18-2006/.

Hao Fang, Jiawei Kong, Wenbo Yu, Bin Chen, Jiawei Li, Hao Wu, Shu-Tao Xia, and Ke Xu. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4090–4100, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Fang_One_Perturbation_is_Enough_On_Generating_Universal_Adversarial_Perturbations_against_ICCV_2025_paper.html.

Jiyuan Fu, Zhaoyu Chen, Kaixun Jiang, Haijing Guo, Jiafeng Wang, Shuyong Gao, and Wenqiang Zhang. Improving adversarial transferability of vision-language pre-training models through collaborative multimodal interaction. *arXiv preprint arXiv:2403.10883*, 2024. doi: 10.48550/arXiv.2403.10883. URL https://arxiv.org/abs/2403.10883v2.

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 *Computer Vision – ECCV 2024*, volume 15115 of *Lecture Notes in Computer Science*, pp. 442–460. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72998-0_25. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/7488_ECCV_2024_paper.php.

Siddhant Garg and Goutham Ramakrishnan. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6174–6181. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.498. URL https://aclanthology.org/2020.emnlp-main.498/.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6572.

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. doi: 10.48550/arXiv.2312.04913. URL https://arxiv.org/abs/2312.04913v1.

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. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8018–8025, 2020. doi: 10.1609/aaai.v34i05.6311. URL https://ojs.aaai.org/index.php/AAAI/article/view/6311.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015. doi: 10.1109/CVPR.2015.7298932. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.html.

Valentin Khrulkov and Ivan Oseledets. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8562–8570, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Khrulkov_Art_of_Singular_CVPR_2018_paper.html.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 2021. URL https://proceedings.mlr.press/v139/kim21k.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

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 *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

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 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6193–6202. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.emnlp-main.500. URL https://aclanthology.org/2020.emnlp-main.500/.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision – ECCV 2020*, volume 12375 of *Lecture Notes in Computer Science*, pp. 121–137. Springer, 2020b. doi: 10.1007/978-3-030-58577-8_8. URL https://link.springer.com/chapter/10.1007/978-3-030-58577-8_8.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

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 *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Xuannan Liu, Yaoyao Zhong, Yuhang Zhang, Lixiong Qin, and Weihong Deng. Enhancing generalization of universal adversarial perturbation through gradient aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4435–4444, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Liu_Enhancing_Generalization_of_Universal_Adversarial_Perturbation_through_Gradient_Aggregation_ICCV_2023_paper.html.

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. URL https://openaccess.thecvf.com/content/ICCV2023/html/Lu_Set-level_Guidance_Attack_Boosting_Adversarial_Transferability_of_Vision-Language_Pre-training_Models_ICCV_2023_paper.html.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1706.06083.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.html.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765–1773, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Moosavi-Dezfooli_Universal_Adversarial_Perturbations_CVPR_2017_paper.html.

Konda Reddy Mopuri, Utsav Garg, and R. Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. In *Proceedings of the British Machine Vision Conference*, 2017. URL https://arxiv.org/abs/1707.05572.

Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R. Venkatesh Babu. NAG: Network for adversary generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 742–751, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Mopuri_NAG_Network_for_CVPR_2018_paper.html.

Konda Reddy Mopuri, Aditya Ganeshan, and R. Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2452–2465, 2019. doi: 10.1109/TPAMI.2018.2861800. URL https://doi.org/10.1109/TPAMI.2018.2861800.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4422–4431, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Poursaeed_Generative_Adversarial_Perturbations_CVPR_2018_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. URL https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Singh_FLAVA_A_Foundational_Language_and_Vision_Alignment_Model_CVPR_2022_paper.html.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514/.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.html.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–2162. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1221. URL https://aclanthology.org/D19-1221/.

Donghua Wang, Wen Yao, Tingsong Jiang, and Xiaoqian Chen. Improving transferability of universal adversarial perturbation with feature disruption. *IEEE Transactions on Image Processing*, 33:722–737, 2024a. doi: 10.1109/TIP.2023.3345136. URL https://doi.org/10.1109/TIP.2023.3345136.

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 *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1722–1740. IEEE, 2024b. doi: 10.1109/SP54263.2024.00102. URL https://doi.org/10.1109/SP54263.2024.00102.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186, 2023a. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_Image_as_a_Foreign_Language_BEiT_Pretraining_for_Vision_and_CVPR_2023_paper.html.

Youze Wang, Wenbo Hu, Yinpeng Dong, and Richang Hong. Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning. *arXiv preprint arXiv:2308.12636*, 2023b. doi: 10.48550/arXiv.2308.12636. URL https://arxiv.org/abs/2308.12636v1. Version 1.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Xie_Improving_Transferability_of_Adversarial_Examples_With_Input_Diversity_CVPR_2019_paper.html.

Fan Yang, Yihao Huang, Kailong Wang, Ling Shi, Geguang Pu, Yang Liu, and Haoyu Wang. Efficient and effective universal adversarial attack against vision-language pre-training models. *arXiv preprint arXiv:2410.11639*, 2024. doi: 10.48550/arXiv.2410.11639. URL https://arxiv.org/abs/2410.11639v2.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15671–15680, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yang_Vision-Language_Pre-Training_With_Triple_Contrastive_Learning_CVPR_2022_paper.html.

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 *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cpDhcsEDC2.

Jiaxin Ye, Weihai Li, Ying Wang, Simeng Qin, Zhitao Zeng, and Zikai Xu. Asymmetric relational-geometry driven universal adversarial perturbations for vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7997–8014, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.393. URL https://aclanthology.org/2026.findings-acl.393/.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a5e3cf29c269b041ccd644b6beaf5c42-Abstract-Conference.html.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, volume 9906 of *Lecture Notes in Computer Science*, pp. 69–85. Springer, 2016. doi: 10.1007/978-3-319-46475-6_5. URL https://link.springer.com/chapter/10.1007/978-3-319-46475-6_5.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25994–26009. PMLR, 2022. URL https://proceedings.mlr.press/v162/zeng22c.html.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5005–5013, 2022. URL https://arxiv.org/abs/2206.09391.

Peng-Fei Zhang and Zi Huang. Hierarchical refinement of universal multimodal attacks on vision-language models, 2026. URL https://arxiv.org/abs/2601.10313v1. arXiv:2601.10313v1, January 15, 2026.

Peng-Fei Zhang, Zi Huang, and Guangdong Bai. Universal adversarial perturbations for vision-language pre-trained models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 862–871, 2024. doi: 10.1145/3626772.3657781. URL https://arxiv.org/abs/2405.05524.

Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. AdvCLIP: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6311–6320, 2023. URL https://arxiv.org/abs/2308.07026.