Vision-fused Attack: Advancing Aggressive and Stealthy Adversarial Text against Neural Machine Translation

Yanni Xue[1*], **Haojie Hao**[1*], **Jiakai Wang**[2†], **Qiang Sheng**[4], **Renshuai Tao**[5], **Yu Liang**[6], **Pu Feng**[1] and **Xianglong Liu**[1,2,3]

[1]State Key Laboratory of Complex & Critical Software Environment, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China
[3]Institute of Data Space, Hefei Comprehensive National Science Center, Anhui, China
[4]Institute of Computing Technology, Chinese Academy of Sciences
[5]Beijing Jiaotong University
[6]Beijing University of Technology
{ynxue, haojiehao,fengpu,xlliu}@buaa.edu.cn, wangjk@mail.zgclab.edu.cn, shengqiang18z@ict.ac.cn, rstao@bjtu.edu.cn, yuliang@bjut.edu.cn

Abstract

While neural machine translation (NMT) models achieve success in our daily lives, they show vulnerability to adversarial attacks. Despite being harmful, these attacks also offer benefits for interpreting and enhancing NMT models, thus drawing increased research attention. However, existing studies on adversarial attacks are insufficient in both attacking ability and human imperceptibility due to their sole focus on the scope of language. This paper proposes a novel vision-fused attack (VFA) framework to acquire powerful adversarial text, *i.e.*, more aggressive and stealthy. Regarding the attacking ability, we design the vision-merged solution space enhancement strategy to enlarge the limited semantic solution space, which enables us to search for adversarial candidates with higher attacking ability. For human imperceptibility, we propose the perception-retained adversarial text selection strategy to align the human text-reading mechanism. Thus, the finally selected adversarial text could be more deceptive. Extensive experiments on various models, including large language models (LLMs) like **LLaMA** and **GPT-3.5**, strongly support that VFA outperforms the comparisons by large margins (up to **81%/14%** improvements on ASR/SSIM).

1 Introduction

Neural machine translation (NMT) has achieved remarkable progress and has been widely used in many scenarios with the advancement of deep neural networks. However, recent studies have revealed the vulnerability of NMT models. A well-designed adversarial text, which aims to deceive NMT models while remaining imperceptible to humans, could result in poor model performance [Zhang *et al.*, 2021]. Nevertheless, as the old saying goes, *every coin has two sides*, though harming NMT models, the adversarial attacks could also help understand the behavior of the unexplainable deep models. Therefore, generating adversarial text has essential value in constructing trustworthy and robust NMT models.

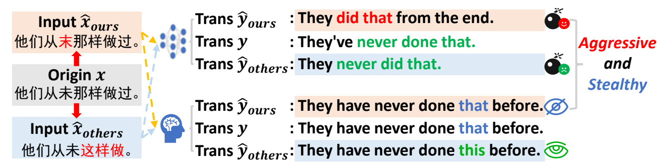

Figure 1: Our proposed VFA reduces the translation quality of NMT models by generating visually similar characters that are aggressive, such as “末” and “未” in the figure, while maintaining consistency with human's recognition of text to achieve stealthiness.

Existing studies on generating adversarial text for NMT models can be classified into targeted attacks and untargeted attacks based on the intended attack purpose. Although both of them aim to fool NMT models, their purposes have a few differences. To be specific, the former simply misleads the translation into arbitrary wrong output, while the latter is intended to let the NMT models make particular false responses, e.g., inserting keywords into the translated sentences [Cheng *et al.*, 2018; Sadrizadeh *et al.*, 2023].

Despite significant progress in generating adversarial text, current studies still show some weaknesses, which can be summarized into two aspects: (1) The unsatisfactory attacking ability. Most attacking methods first search candidate adversarial words in the embedding space, which is strictly constrained due to the extra semantic-preserving requirements. As a result, the searched adversarial results could not mislead the NMT models to the largest extent in such a narrow embedding space. (2) The insufficient imperceptibility goal. Current attacks mainly achieve the goal via restricting the adversarial results at the semantic level. However, the text's perception of mankind is more correlated to the visual system, i.e., humans always recognize text first and process them later. Thus, due

Codes can be found at https://github.com/Levelower/VFA.

to the sole semantic restriction, the generated adversarial text might not be stealthy enough for humans.

To tackle these issues, we propose Vision-fused Attack framework (VFA) to generate more aggressive adversarial text against NMT models with higher attacking ability and better visual stealthiness. Figure 1 illustrates the difference between our adversarial text and others. **To improve the attacking ability**, considering that the limited semantic solution space might restrict the adversarial text candidates, we propose the vision-merged solution space enhancement (VSSE) strategy to abound the searchable adversarial candidates by the visual solution space mixture module. In detail, we first enhance the basic semantic space with the help of a reverse translation block. Further, we map the enhanced solution space into vision space via the text-image transformation block. Since the mapped visual solution space mixes both semantic and visual characteristics, it offers a broader searching range for candidate adversarial words, thus making it more possible to activate higher attacking ability in practice. **Regarding the imperceptibility of adversarial text**, given the neglected fact that human reading accepts visual signals first to recognize text, we develop the perception-retained adversarial text selection (PATS) strategy to evade human perception through the perception stealthiness enhancement module. Specifically, an improved word replacement operation is preliminarily introduced to disperse attack locations. Then, we integrate the visual characteristics of local characters and global sentences to align with the human text-reading mechanism. Since this selection strategy could filter the human perceptually suspected candidates, we could efficiently and accurately select more imperceptible adversarial text in principle to deceive the text perception of humans.

To demonstrate the effectiveness of the proposed method, we conduct extensive experiments under white-box and black-box settings on various representative models and widely-used datasets, including open-source and closed-source large language models like GPT-3.5 and LLaMA. The experimental results strongly support that our **VFA** outperforms the comparisons by large margins.

Our main contributions are as follows:

- To the best of our knowledge, we are the first to introduce visual perception, which aligns with human reading, to generate adversarial text against NMT models.
- We propose a Vision-fused Attack (VFA) against NMT models, acquiring aggressive and stealthy adversarial text through vision-merged solution space enhancement and perception-retained adversarial text selection.
- Extensive experiments show that VFA outperforms the comparisons by large margins (up to 81%/28% improvements on common NMT models/LLMs), and achieves considerable imperceptibility (up to 14% improvements) in both machine evaluation and human study.

2 Related Work

The neural machine translation (NMT), which translates texts from one language to another, has achieved impressive progress using DNN models such as transformers [Vaswani *et al.*, 2017]. Due to their excellent performance, NMT models are widely used in different applications [Gao *et al.*, 2023]. However, recent studies of adversarial text investigate the vulnerability of NMT models, whose ultimate goal is to mislead NMT models but imperceptible to humans. Though they are harmful to DNN models, they could also help us understand them [Wang *et al.*, 2021a; Wang *et al.*, 2021b]. Therefore, there has been an increasing number of recent studies on adversarial attacks against NMT models.

Adversarial attacks on NMT models can be divided into untargeted attacks and targeted attacks. Untargeted attacks aim to degrade the translation quality of NMT models. Targeted attacks mislead the victim model to produce a particular translation, for example, one that does not overlap with the reference or inserts some keywords into the translation.

Current untargeted attacks always use semantic perturbations to degrade the model performance. Specifically, attackers rank the words in a sentence by saliency and randomly substitute them based on the saliency order [Cheng *et al.*, 2019; Cheng *et al.*, 2020]. However, random substitution may reduce the imperceptibility of adversarial text. Therefore, some studies choose to replace these words with similar ones, requiring semantic similarity techniques to identify similar word lists [Feng *et al.*, 2022]. Nevertheless, the inefficiency of embedding techniques may still result in false substitutions. To resolve this problem, Word Saliency speedup Local Search uses Round-Trip Translation (RTT) to handle it [Zhang *et al.*, 2021]. Additionally, the Doubly Round-Trip Translation (DRTT) method further amplifies the RTT to achieve better quality of adversarial text [Lai *et al.*, 2022].

In targeted attack research, attacks fool the target model into generating a particular translation. Such as the translation does not overlap with the reference or push some words into it [Cheng *et al.*, 2018]. They use a hinge-like loss term and a group lasso regularization to make perturbations, thereby achieving specific attack targets. Since these perturbations in the embedding space are less constrained, they may not preserve semantic similarity. Therefore, some studies define the new optimization problem, including adversarial loss and similarity terms. Finally, they perform gradient projection in the embedding space to generate adversarial sentences [Sadrizadeh *et al.*, 2023].

Both above approaches aim to perturb original sentences with similar words, which satisfies the requirement of semantics-preserving. However, there are differences in the cognitive process of text between humans and models. For humans, the cognitive process of text can be divided into visual perception and semantic understanding, while models only imitate the latter. Therefore, the generated adversarial text does not perform well in the visual perception part, resulting in poor human imperceptibility.

3 Approach

In this section, we define the adversarial text for NMT models and then elaborate on our proposed Vision-fused Attack.

3.1 Problem Definition

Given a source sentence $\mathbf{x}$ and a reference sentence $\mathbf{y}$, the adversarial text $\mathbf{x}_\delta$ is generated to mislead targeted model

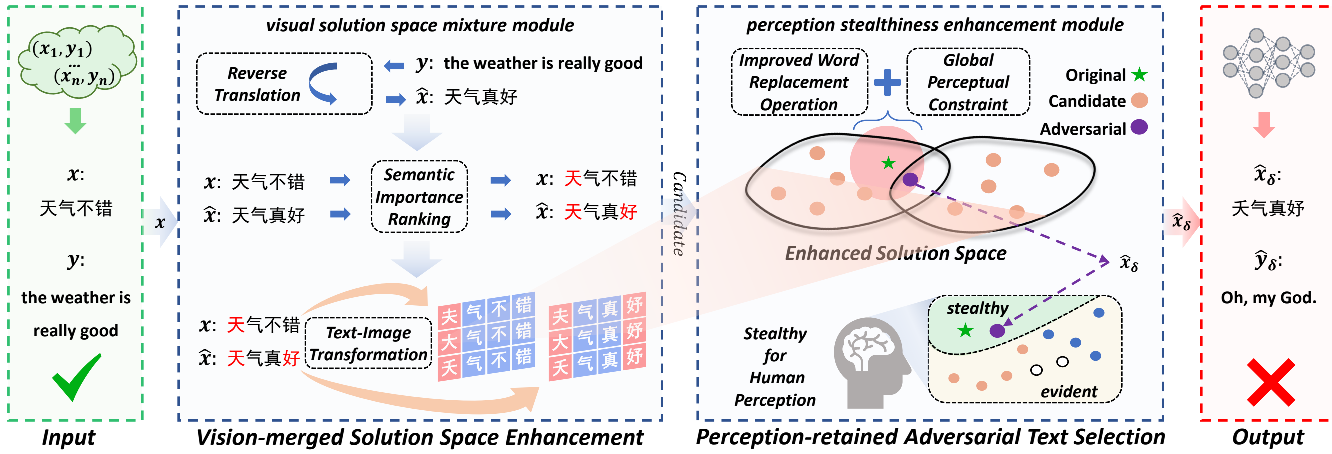

Figure 2: Overall framework of our Vision-fused Attack (VFA). We first use the Vision-merged Solution Space Enhancement strategy to search for more aggressive adversarial candidates. Then we find the final adversarial text that best matches human perception through the Perception-retained Adversarial Text Selection strategy. Finally we find more aggressive and stealthy adversarial text against NMT models.

$\mathbb{M}$ to generate low-quality results. Referring to previous work [Ebrahimi *et al.*, 2018a], a successful attack satisfies:

$$\frac{\mathrm{sim_t}(\mathbb{M}(\mathbf{x}), \mathbf{y}) - \mathrm{sim_t}(\mathbb{M}(\mathbf{x}_\delta), \mathbf{y})}{\mathrm{sim_t}(\mathbb{M}(\mathbf{x}), \mathbf{y})} > \alpha. \tag{1}$$

Here, $\mathrm{sim_t}(\cdot)$ refers to the similarity function used for evaluating semantic similarity. The parameter α is established as the lower limit, representing the lowest degradation of translation quality. It is worth noting that we differ from previous work in assessing effectiveness, which uses sole semantic similarity [Zhang *et al.*, 2021; Lai *et al.*, 2022]. We define an authentic adversarial text from the perspective of visual perception, and the definition of adversarial text differs accordingly. Specifically, we generate a semantically extension $\hat{\mathbf{x}}$ for the original text $\mathbf{x}$ (this process will be detailed in the Reverse Translation Block), and use $\hat{\mathbf{x}}$ as the text to be replaced to generate the final adversarial text $\mathbf{x}_\delta$. Our modified definition is given as:

$$\begin{cases} \dfrac{\mathrm{sim_t}(\mathbb{M}(\mathbf{x}), \mathbf{y}) - \mathrm{sim_t}(\mathbb{M}(\mathbf{x}_\delta), \mathbf{y})}{\mathrm{sim_t}(\mathbb{M}(\mathbf{x}), \mathbf{y})} > \alpha, \\ \mathrm{sim_t}(\mathbf{x}, \hat{\mathbf{x}}) > \beta, \\ \mathrm{sim_v}(\mathbf{x}, \mathbf{x}_\delta) > \theta. \end{cases} \tag{2}$$

The function $\mathrm{sim_v}$ is a visual similarity function defined by the perception stealthiness enhancement module, which will be detailed in the following subsection. The parameter β and θ ensure the semantic and visual similarity. The authentic adversarial text satisfies the requirements of visual and semantic similarity while degrading the translation quality.

3.2 Overview of Vision-fused Attack

Previous studies generated adversarial text within a limited semantic space and overlooked the significance of visual perception in text reading. Consequently, it is challenging to produce more aggressive and imperceptible results. In response, we introduce a Vision-fused Attack (VFA) to generate human imperceptible adversarial text against NMT models. The overall framework is illustrated in Figure 2.

Addressing the limitation of the semantic solution space, we consider the amplified visual solution space and generate effective adversarial text through the Visual-merged Solution Space Enhancement (VSSE) strategy. Specifically, we expand the essential semantic solution space using a reverse translation block. Furthermore, our method generates candidate adversarial words through a text-image transformation block. With abundant adversarial candidates, we then filter unauthentic candidates using the Perception-retained Adversarial Text Selection (PATS) strategy to acquire more imperceptible adversarial text. Initially, we perform substitutions through an improved word replacement operation. Subsequently, we obtain the authentic adversarial text with a global perceptual constraint. This enables us to achieve superior perceptual retention results, aligning more closely with vision-correlated human text perception.

3.3 Vision-merged Solution Space Enhancement

Our vision-merged enhanced solution space consists of two parts: one is the essential semantic space expanded through reverse translation block, which is termed as $\mathbb{T}$, and the other is the mapped visual solution space corresponding to the input text, which is termed as $\mathbb{V}$. Considering the further enhancement of vision-merged solution space, we transform the original input to increase the variety of words. Therefore, we can map semantic space into a larger visual space and search for adversarial words within this expanded area. When generating the candidate adversarial words, we first tokenize the original input sentence and the texts which are transformed through the reverse translation block. Then, we use semantic importance ranking to obtain ordered attack locations. Finally, we acquire possible adversarial candidates through a text-image transformation block.

Reverse Translation Block. We initially expand essential semantic space through reverse translation to amplify vision-fused solution space. For the source sentence $\mathbf{x}$ and the reference sentence $\mathbf{y}$, we generate similar sentences (referred to as $\hat{\mathbf{x}}$) for $\mathbf{x}$. For each reference translation $\mathbf{y}$, we use an auxiliary translation model $\mathbf{M}_{\text{aux}}$ to obtain transformed $\hat{\mathbf{x}}$ meanwhile the similarity between $\mathbf{x}$ and $\hat{\mathbf{x}}$ need to satisfy the constraint of the lowest thresh β. The sentence similarity is evaluated by a multilingual sentence model, which is termed as $\mathbf{M}_{\text{sim}}$:

$$\hat{\mathbf{x}} = \mathbf{M}_{\text{aux}}(\mathbf{x}),\ \ \mathbf{M}_{\text{sim}}(\hat{\mathbf{x}}, \mathbf{x}) > \beta. \tag{3}$$

Semantic Importance Ranking. The source sentence can be represented as a list of words $\mathbf{w} = \langle w_1, w_2, ..., w_n\rangle$. For a sequence that is masked at position i, which is termed as $\mathbf{w}^{\mathbf{i}}_{\text{mask}} = \langle w_1, w_2, ..., w_{i-1}, [m], w_{i+1}, ..., w_n\rangle$. We calculate the importance scores of these words at different positions named $\mathbf{w}^{i}_{\text{imp}}$. Masked language model $\mathbf{M}_{\text{mlm}}$ (e.g., BERT) is used to predict word probability of occurrence. This process can be written as:

$$\mathbf{w}^{i}_{\text{imp}} = \mathbf{M}_{\text{mlm}}(w_i|\langle w_1, ..., w_{i-1}, [m], w_{i+1}, ..., w_n\rangle). \tag{4}$$

Text-image Transformation Block. For each character c in the Unicode character dictionary (referred to as $\mathbb{C}$), we conduct a level-wise search to find the most similar candidate adversarial characters. Initially, we follow previous work to utilize the glyph dictionary, denoted as $\mathbb{D}$, which stores mappings of specific characters c and their radicals z (Here we denote the set of all radicals as $\mathbb{Z}$, where $c \in \mathbb{C}$ and $z \in \mathbb{Z}$) [Su *et al.*, 2022]. Each radical of character c can be found through function $f_c(\cdot)$. We start by aggregating characters with the same radical (denoted as c'), narrowing down the candidate pool. We can obtain a portion of the candidate set $\mathbf{S}_{\text{rad}}$ through a similar radical search, denoted as $f_a(\cdot)$:

$$\begin{cases} f_c : c \rightarrow \{z \mid z \in \mathbb{Z}, (c, z) \in \mathbb{D}\}, \\ f_a : c \rightarrow \{c' \mid f_c(c) \cap f_c(c') \neq \emptyset\}. \end{cases} \tag{5}$$

Due to the limitations of the glyph dictionary, we further pixelated the characters set and conducted similar image searches to find similar characters, which is denoted as $\mathbf{S}_{\text{pix}}$. The function $\mathrm{p}(\cdot)$ is defined to convert a sentence or character into the correlated image. We map the input character to its pixelated image through $\mathrm{p}(c)$. Then, we calculate cosine similarity to search $top\ m$ visually similar results using Faiss (a tool that accelerates vector calculations through hierarchical search) [Johnson *et al.*, 2019]. And the procedure could be formulated as $f_{cos}(\cdot)$:

$$f_{\text{cos}} : c \rightarrow \{top(m, cos(\mathrm{p}(c), \mathrm{p}(c'))), c' \in \mathbb{C}\}, \tag{6}$$

where the $top(\cdot)$ represents the function that identifies the highest-ranked elements based on certain scores, such as cosine similarity. Finally, we apply Mean Squared Error (MSE) similarity to re-rank the possible adversarial character and select the $top\ k$ candidate result:

$$\mathbf{S}_{\text{pix}} = top(k, \text{mse}(f_{\text{cos}}(c), c)). \tag{7}$$

Combined with the results of $\mathbf{S}_{\text{rad}}$ and $\mathbf{S}_{\text{pix}}$, we can obtain the final candidate adversarial characters set $\mathbf{S}$.

$$\mathbf{S} = \mathbf{S}_{\text{rad}} \cup \mathbf{S}_{\text{pix}}. \tag{8}$$

3.4 Perception-retained Adversarial Text Selection

For the generated candidate characters set $\mathbf{S}$, we apply a perception stealthiness enhancement module, which consists of improved word replacement operation and global perceptual constraint, to filter out truly effective adversarial text. Therefore, the visually imperceptible candidates could stand out to better mislead human perception, *i.e.*, stealthy.

Improved Word Replacement Operation. Firstly, we implement a substitution constraint strategy grounded in human perception. We intuitively regulate the replacement rate, deliberately replacing only one character within each word. This strategy disrupts the semantic expression of the word and minimizes the impact of perturbation. For text $\hat{\mathbf{x}}$ to be replaced, w_i denotes the i-th word in order of importance $\mathbf{w}^{i}_{\text{imp}}$ of the text, and c_j signifies the j-th character of a word w_i. The replacement operation rate, denoted as r, falls within the range $0 \leq r \leq 1$, representing the replacement probability of the overall sentence. And function $\text{rep}(c_{j_\delta}, c_j)$ indicates whether the character c_j is substituted by c_{j_δ} in $\mathbf{S}$. When the character changes, the function equals 1. Otherwise, it equals 0. We can detail this constraint as follows:

$$\begin{cases} \dfrac{\sum_{w_i \in \mathbf{x}_\delta} \sum_{c_j \in w_i} \text{rep}(c_{j_\delta}, c_j)}{\sum_{w_i \in \hat{\mathbf{x}}} \sum_{c_j \in w_i} 1} < r, \\ \forall w_i \in \mathbf{x}_\delta, \quad \displaystyle\sum_{c_j \in w_i} \text{rep}(c_{j_\delta}, c_j) \leq 1. \end{cases} \tag{9}$$

Global Perceptual Constraint. Moreover, we take visual perception constraints into account. We introduce a visual perceptual similarity score for batch assessment of visual similarity. For a source sentence $\mathbf{x}$ paired with a reference image represented as $\mathrm{p}(\mathbf{x})$, and a perturbed sentence $\mathbf{x}_\delta$ with its associated image denoted as $\mathrm{p}(\mathbf{x}_\delta)$, $\mathbb{L}(a, b)$ denotes the perceptual similarity score between images a and b. The parameter ϵ serves as a weight for local perception, where $0 \leq \epsilon \leq 1$. The sentence visual similarity score can be calculated using the visual perception constraint strategy. We introduce the LPIPS metric to construct a global perceptual constraint [Zhang *et al.*, 2018]. This constraint measures the global perceptual similarity of a sentence and aggregates the local perceptual similarities of characters with the weighted summation. Finally, we use the perceptual constraint threshold θ to constrain visual similarity:

$$\mathbb{L}(\mathrm{p}(\mathbf{x}_\delta), \mathrm{p}(\mathbf{x})) + \epsilon \sum_{w_i \in \mathbf{x}_\delta} \sum_{c_j \in w_i} \mathbb{L}(\mathrm{p}(c_{j_\delta}), \mathrm{p}(c_j)) > \theta. \tag{10}$$

4 Experiment

In this section, we first describe the experimental settings, and then we report the experimental results and some discussions on the common NMT models and LLMs.

4.1 Experiments Settings

Datasets and Models. We choose the validation set of WMT19 [Ng *et al.*, 2019], WMT18 [Bojar *et al.*, 2018], and TED [Cettolo *et al.*, 2016] for the Chinese-English (Zh-En) translation task and the test set of ASPEC [Nakazawa *et al.*, 2016] for the Japanese-English (Ja-En) translation task. Regarding the models, the NMT models for both translation tasks are implemented using HuggingFace's Marian Model [Junczys-Dowmunt *et al.*, 2018], with the Zh-En/Ja-En translation models as the targeted models and the En-Zh/En-Ja models as the auxiliary models. These datasets and models are widely used in previous studies. Besides, we consider pixel-based machine translation model as the targeted model to test the validity of our method [Salesky *et al.*, 2023].

Method	WMT19 (Zh-En)			WMT18 (Zh-En)			TED (Zh-En)			ASPEC (Ja-En)		
	BLEU↓	ASR↑	SSIM↑	BLEU↓	ASR↑	SSIM↑	BLEU↓	ASR↑	SSIM↑	BLEU↓	ASR↑	SSIM↑
Baseline	0.178			0.163			0.159			0.075		
HotFlip	$0.141_{\downarrow 21\%}$	0.213	0.717	$0.131_{\downarrow 19\%}$	0.212	0.722	$0.121_{\downarrow 24\%}$	0.208	0.737	$0.047_{\downarrow 38\%}$	0.334	0.717
Seq2Sick	$0.139_{\downarrow 22\%}$	0.198	0.773	$0.134_{\downarrow 18\%}$	0.164	0.777	$0.112_{\downarrow 30\%}$	0.228	0.762	$0.072_{\downarrow 4\%}$	0.030	**0.862**
Targeted	$0.134_{\downarrow 25\%}$	0.234	0.799	$0.126_{\downarrow 22\%}$	0.211	0.801	$0.114_{\downarrow 28\%}$	0.266	0.793	$0.047_{\downarrow 37\%}$	0.308	0.734
DRTT	$0.144_{\downarrow 19\%}$	0.173	0.768	$0.131_{\downarrow 19\%}$	0.173	0.760	$0.136_{\downarrow 14\%}$	0.130	0.780	$0.069_{\downarrow 8\%}$	0.063	0.850
ADV	$0.146_{\downarrow 18\%}$	0.174	0.838	$0.142_{\downarrow 12\%}$	0.141	0.843	$0.125_{\downarrow 21\%}$	0.200	0.842	-	-	-
Ours	$\mathbf{0.107_{\downarrow 40\%}}$	**0.382**	**0.950**	$\mathbf{0.097_{\downarrow 40\%}}$	**0.384**	**0.949**	$\mathbf{0.109_{\downarrow 31\%}}$	**0.299**	**0.964**	$\mathbf{0.042_{\downarrow 44\%}}$	**0.387**	0.859

Table 1: Performance of VFA on Zh-En and Ja-En translation tasks using pure-text NMT models. The "Baseline" row records the metric scores on the clean dataset. The remaining rows record the different metric scores and the decrease relative to the "Baseline" of adversarial texts generated by various methods. ↓ indicates the lower, the better, and ↑ is the opposite.

Evaluation Metrics. We use the relative decrease of the BLEU to measure the aggressiveness of adversarial text [Papineni *et al.*, 2002]. A successful attack is defined when the BLEU score decreases by over 50%. The attack success rate (ASR) is defined as the ratio of successful adversarial texts to the total. Finally, we use the SSIM value to evaluate the imperceptibility of the adversarial text [Wang *et al.*, 2004].

Compared Methods. We choose several state-of-the-art works about NLP attack and NMT attack, including HotFlip [Ebrahimi *et al.*, 2018b], Seq2Sick [Cheng *et al.*, 2018], Targeted Attack [Sadrizadeh *et al.*, 2023], DRTT [Lai *et al.*, 2022] and ADV [Su *et al.*, 2022].

Implementation Details. As for the hyperparameter settings, we set the global perception constraint to 0.95 and the replacement rate to 0.2. To evaluate the semantic similarity between two sentences, we employ the HuggingFace sentence-transformer model [Wang *et al.*, 2020], which supports multiple languages. Additionally, we utilize a Bert architecture model [Cui *et al.*, 2019] to predict the importance of words. We conduct experiments in a cluster of NVIDIA GeForce RTX 3090 GPUs.

4.2 Effectiveness on Common NMT Models

In this section, we evaluate the aggressiveness and imperceptibility of adversarial texts generated by our VFA and compared methods. The evaluation is conducted on the Zh-En and Ja-En translation tasks using pure-text and pixel-based NMT models.

As seen in Tables 1 and 2, **our VFA generates adversarial texts with the highest aggressiveness across both pixel-based and pure-text models in both tasks.**

(1) In terms of aggressiveness, our VFA achieves the maximum BLEU decrease and the highest ASR on different datasets and translation tasks. Taking the results on WMT18 as an example, our VFA achieves a BLEU decrease of 40%, better than the best of 22% achieved by the Targeted Attack. Our ASR (0.384) outperforms the best (0.212) by 81%. This indicates that our proposed Vision-merged Solution Space Enhancement strategy effectively improves aggressiveness.

(2) Regarding imperceptibility, our VFA also achieves the best result. Our VFA maintains an SSIM value above 0.94 on Zh-En translation tasks, while the ADV has an SSIM value of no more than 0.85. Although our VFA is second only to Seq2Sick in the imperceptibility of the Ja-En translation task, Seq2Sick nearly keeps the original sentence when attacking Japanese texts, resulting in lower ASR and inability to guarantee aggressiveness. The results demonstrate that our proposed Perception-retained Adversarial Text Selection strategy effectively improves the imperceptibility of adversarial texts.

Method	WMT19 (Zh-En)		WMT18 (Zh-En)		TED (Zh-En)	
	BLEU↓	ASR↑	BLEU↓	ASR↑	BLEU↓	ASR↑
Baseline	0.069		0.070		0.165	
HotFlip	$0.054_{\downarrow 22\%}$	0.216	$0.057_{\downarrow 18\%}$	0.201	$0.130_{\downarrow 21\%}$	0.202
Seq2Sick	$0.055_{\downarrow 21\%}$	0.209	$0.058_{\downarrow 17\%}$	0.189	$0.134_{\downarrow 18\%}$	0.183
Targeted	$0.056_{\downarrow 19\%}$	0.162	$0.060_{\downarrow 15\%}$	0.154	$0.143_{\downarrow 13\%}$	0.105
DRTT	$0.058_{\downarrow 16\%}$	0.139	$0.060_{\downarrow 14\%}$	0.125	$0.140_{\downarrow 15\%}$	0.116
ADV	$0.062_{\downarrow 10\%}$	0.133	$0.063_{\downarrow 10\%}$	0.123	$0.138_{\downarrow 16\%}$	0.149
Ours	$\mathbf{0.053_{\downarrow 23\%}}$	**0.232**	$\mathbf{0.056_{\downarrow 20\%}}$	**0.220**	$\mathbf{0.128_{\downarrow 23\%}}$	**0.215**

Table 2: Performance of VFA on the pixel-based NMT model.

4.3 Transferability on LLMs

In this section, we evaluate the transferability of our VFA through LLM testing. We use adversarial texts generated by the common NMT model to evaluate the BLEU decrease and ASR on LLM. Our selection includes four models: LLaMA-13B [Touvron *et al.*, 2023], BaiChuan-13B [Baichuan, 2023], GPT-3.5-turbo [OpenAI, 2022], and Wenxin Yiyan (ERNIE) [Baidu, 2023], representing leading models in both open-source and closed-source fileds. ChatGPT and LLaMA represent the most advanced LLM for English, while ERNIE and BaiChuan represent the most advanced LLM for Chinese. These LLMs perform much better on clean datasets than common NMT models, indicating that LLMs also have strong capabilities in translation tasks.

Table 3 displays the experiments on these models. These results indicate that **even large language models exhibit decreased performance in the face of attack.** We further give some insights and discussions as follows:

(1) Our VFA demonstrates the best aggressiveness in the open-source and closed-source LLMs. For the widely used LLaMA and ChatGPT, our VFA achieves the strongest aggressiveness on all datasets. For BaiChuan and ERNIE, which perform better in Chinese, our VFA also performs best on WMT19 and WMT18. This proves that our VFA has good transferability for LLMs.

Method	Model	WMT19		WMT18		TED		Model	WMT19	
		BLEU↓	ASR↑	BLEU↓	ASR↑	BLEU↓	ASR↑		BLEU↓	ASR↑
Baseline	LLaMA (open)	0.156		0.146		0.113		ChatGPT (closed)	0.226	
HotFlip		$0.119_{\downarrow 23\%}$	0.289	$0.113_{\downarrow 23\%}$	0.249	$0.083_{\downarrow 27\%}$	0.273		$0.188_{\downarrow 17\%}$	0.205
Seq2Sick		$0.118_{\downarrow 24\%}$	0.273	$0.113_{\downarrow 23\%}$	0.241	$0.073_{\downarrow 35\%}$	0.307		$0.180_{\downarrow 20\%}$	0.235
Targeted		$0.115_{\downarrow 26\%}$	0.306	$0.109_{\downarrow 25\%}$	0.281	$0.081_{\downarrow 29\%}$	0.298		$0.175_{\downarrow 23\%}$	0.237
DRTT		$0.126_{\downarrow 19\%}$	0.233	$0.119_{\downarrow 18\%}$	0.224	$0.098_{\downarrow 14\%}$	0.166		$0.181_{\downarrow 20\%}$	0.231
ADV		$0.132_{\downarrow 15\%}$	0.214	$0.128_{\downarrow 12\%}$	0.174	$0.093_{\downarrow 18\%}$	0.213		$0.200_{\downarrow 12\%}$	0.156
Ours		$\mathbf{0.099}_{\downarrow \mathbf{37\%}}$	**0.385**	$\mathbf{0.094}_{\downarrow \mathbf{35\%}}$	**0.359**	$\mathbf{0.071}_{\downarrow \mathbf{37\%}}$	**0.325**		$\mathbf{0.168}_{\downarrow \mathbf{26\%}}$	**0.281**
Baseline	BaiChuan (open)	0.227		0.199		0.145		ERNIE (closed)	0.275	
HotFlip		$0.187_{\downarrow 17\%}$	0.182	$0.167_{\downarrow 17\%}$	0.155	$0.113_{\downarrow 22\%}$	0.201		$0.223_{\downarrow 19\%}$	0.157
Seq2Sick		$0.185_{\downarrow 18\%}$	0.169	$0.167_{\downarrow 17\%}$	0.171	$\mathbf{0.105}_{\downarrow \mathbf{27\%}}$	0.215		$0.213_{\downarrow 22\%}$	0.198
Targeted		$0.179_{\downarrow 21\%}$	0.199	$0.159_{\downarrow 20\%}$	0.203	$0.107_{\downarrow 26\%}$	**0.228**		$0.209_{\downarrow 24\%}$	0.213
DRTT		$0.182_{\downarrow 20\%}$	0.183	$0.161_{\downarrow 19\%}$	0.199	$0.126_{\downarrow 13\%}$	0.133		$0.220_{\downarrow 20\%}$	0.187
ADV		$0.206_{\downarrow 9\%}$	0.110	$0.183_{\downarrow 8\%}$	0.110	$0.122_{\downarrow 16\%}$	0.143		$0.244_{\downarrow 11\%}$	0.113
Ours		$\mathbf{0.176}_{\downarrow \mathbf{23\%}}$	**0.211**	$\mathbf{0.158}_{\downarrow \mathbf{21\%}}$	**0.209**	$0.110_{\downarrow 24\%}$	0.209		$\mathbf{0.205}_{\downarrow \mathbf{25\%}}$	**0.223**

Table 3: Performance of our VFA on open-source and closed-source LLMs, respectively. Our VFA achieves considerable attacking ability.

(2) An interesting phenomenon is that our VFA is more aggressive against English LLMs than other methods. However, when applied to Chinese LLMs, our VFA only achieved a slight lead in aggressiveness and even performed weaker than Targeted Attack and Seq2Sick on the TED dataset. We attribute this phenomenon to the learning of human perception by LLM. After training with a large amount of Chinese corpus, the large model can generalize to a certain degree of human-like perception ability, which makes them robust to visual adversarial texts generated by our VFA.

4.4 Human Study

To evaluate the impact of adversarial texts generated by our VFA and compared methods on reading comprehension, we conduct a human perception study on SurveyPlus, which is one of the most commonly used crowdsourcing platforms.

We select 20 adversarial texts generated by six methods that meet the definition of a successful attack, then we select 105 subjects and conduct the following human perception experiments: (1) **Semantic understanding.** Subjects are informed of some changes in each text and asked if these changes affect their understanding of semantics. (2) **Semantic comparison.** For the texts selected by the subjects in the previous stage, they are required to compare them with original texts and choose the ones that maintain semantic consistency. Finally, we count the number of texts in each method that do not affect understanding (score1) and are semantically consistent with the original texts (score2), and divide them by the total number of people as the scores for each method.

From Table 4, it can be seen that our VFA achieves the second-highest score in the semantic understanding stage and the highest score in the semantic comparison stage. This indicates that our method has the least impact on human understanding and the least impact on semantic changes. As for DRTT, which achieved the highest score in semantic understanding, due to its use of grammatically similar modification strategies, it will have a significant change in the semantics of the original sentence, resulting in a lower score in the semantic comparison stage. In the total score of the two stages, VFA achieves the highest, indicating that it has the least impact on human reading comprehension and has good imperceptibility.

Method	Ours	ADV	DRTT	HotFlip	Targeted	Seq2Sick
score1↑	9.35	6.13	**9.76**	5.07	6.66	5.90
score2↑	**8.63**	5.77	5.49	4.31	4.29	3.88
sum↑	**17.98**	11.90	15.25	9.38	10.95	9.78

Table 4: Result of human study. We add up the scores of the two stages (score1+score2) as the total score (sum) for each method.

4.5 Ablation Study

In this section, we conducted several ablation studies to further investigate the contributions of some crucial components and hyper-parameters in our method.

Effectiveness of Enhanced Solution Space and Perception Constraint. We divide our VFA into Vision-merged Solution Space Enhancement (VSSE) and Perception-retained Adversarial Text Selection (PATS). For the vision-merged solution space, we compare the aggressiveness with the semantic solution space. Additionally, we explore the impact of PATS on the imperceptibility of adversarial texts. We conduct four sets of ablation experiments on these two components, namely, whether to use VSSE and PATS. As shown in Table 5, it is evident that searching in the enhanced solution space significantly improves aggressiveness compared to semantic solution space. Additionally, the PATS ensures a

Ablation	VSSE	PATS	TIT	BLEU↓	ASR↑	SSIM↑
VSSE + PATS	✓	✓	rad+pix	$0.107_{\downarrow 40\%}$	0.382	0.950
	✓	✗	rad+pix	$0.077_{\downarrow 57\%}$	0.542	0.922
	✗	✓	rad+pix	$0.176_{\downarrow 0.8\%}$	0.013	0.993
	✗	✗	rad+pix	$0.117_{\downarrow 34\%}$	0.316	0.662
TIT	✓	✓	rad+pix	$0.107_{\downarrow 40\%}$	0.382	0.950
	✓	✓	pix	$0.092_{\downarrow 49\%}$	0.476	0.944
	✓	✓	rad	$0.125_{\downarrow 30\%}$	0.283	0.955

Table 5: Effectiveness of different components in our VFA.

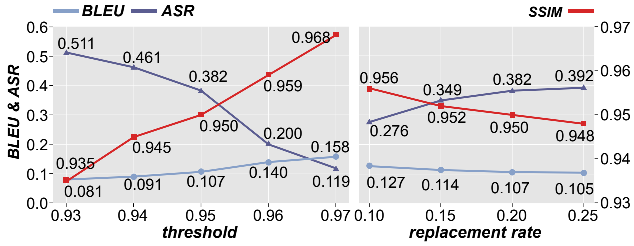

Figure 3: Effectiveness of hyper-parameters.

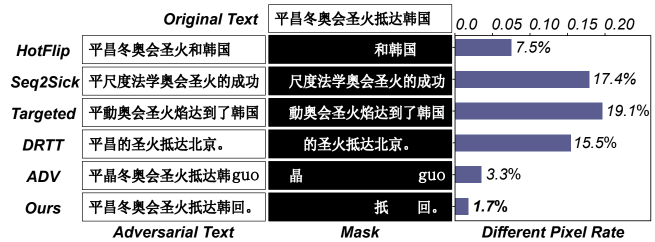

Figure 5: Adversarial texts of all methods and their masks.

higher SSIM similarity score between the original texts and adversarial texts, which means ensuring the imperceptibility of adversarial texts. Meanwhile, adversarial texts generated using only semantic solution space violate the visual imperceptible rule. Therefore, after applying PATS, the ASR of semantic space search will significantly decrease.

Effectiveness of Different Parts in Text-Image Transformation. In the Text-image Transformation (TIT) block, we employed two strategies to form a complete visual solution space, one formed through similar radical components (rad) and the other through pixel characters (pix). Therefore, we explore the impact of these two complementary strategies on aggressiveness and imperceptibility. As shown in Table 5, the results indicate that the visual solution space formed by pixel-based strategies has stronger aggressiveness, while radical-based strategies have stronger imperceptibility. Therefore, by using the former as a supplement to the latter, we ultimately form a solution space that is both aggressive and has extremely high imperceptibility.

Impacts of Hyper-Parameters. We analyze the impact of various visual perception constraint thresholds θ (threshold) and different replacement rates r (rate) of aggressiveness and imperceptibility. As shown in Figure 3, we can draw the following conclusion: (1) It can be observed that the stronger the visual perception constraint, the higher the SSIM, indicating that visual perception constraint can significantly improve the imperceptibility of adversarial texts. (2) It can be observed that a higher replacement rate leads to an increase in ASR but a decrease in SSIM, which is consistent with our expectations. As more words are replaced, the difference between the adversarial texts and the original texts becomes greater, and the aggressiveness also increases. We balanced aggressiveness and imperceptibility by combining the replacement rate and visual perception constraints.

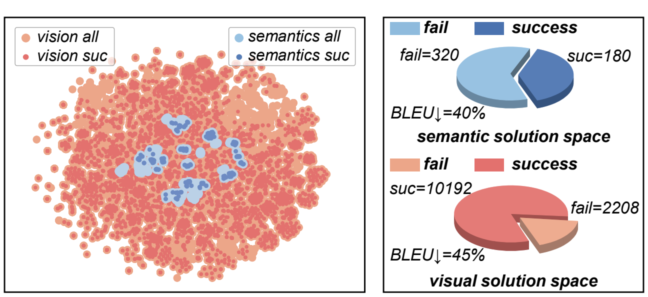

Figure 4: Enhanced solution space and semantic solution space.

4.6 Case Study

Enhanced Solution Space Analysis. We analyze the enhanced solution space from a visualization perspective. For a single text, we calculated embeddings for all texts in the enhanced solution space and the semantic solution space, then reduced the dimensionality using t-SNE [van der Maaten and Hinton, 2008]. The result is shown in Figure 4. The left side of the figure shows the sample distribution in the two solution spaces, while the right side indicates the number of successfully attacked texts and the corresponding BLEU decrease in both spaces. Generally, we can draw such conclusions: (1) It can be witnessed that the scope of the enhanced solution space is broader than that of the semantic solution space. This indicates that the enhanced solution space can obtain more diverse adversarial texts, providing more possibilities to find more effective adversarial texts. (2) The pie charts of both solution spaces show that the enhanced solution space generates more aggressive texts than the semantic solution space. This is further proven by the BLEU decrease of the two approaches. Therefore, an enhanced solution space helps generate more texts with stronger aggressiveness.

Imperceptibility Analysis. We visualized the adversarial texts generated by all methods and analyzed the imperceptibility of individual examples. Figure 5 shows the adversarial texts generated by our method and five compared methods. The differences between them and the original texts are presented in the form of masks. The right side of the figure indicates the proportion of the differences in pixel values between the adversarial texts and the original texts. From the figure, we can see that at the cognitive level, our method requires minimal changes to the original texts, thus achieving recognition consistency and maintaining semantic consistency with the original texts in human reading comprehension.

5 Conclusion

This paper proposed a vision-fused attack (VFA) framework for generating powerful adversarial text. Our VFA uses the vision-merged solution space enhancement and perception-retained adversarial text selection strategy, producing more aggressive and stealthy adversarial text against NMT models. Extensive experiments demonstrated that VFA outperforms comparisons by significant margins both in attacking ability and imperceptibility enhancements. This paper indicates the important effect of multimodal correlations in current deep learning, which encourages future investigations on the corresponding topics, *e.g.*, adversarial defense.

Acknowledgements

This work was supported by Grant KZ46009501.

Contribution Statement

Yanni Xue and Haojie Hao contributed equally to this work. Jiakai Wang is the corresponding author.

References

[Baichuan, 2023] Baichuan. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

[Baidu, 2023] Baidu. Introducing ernie 3.5: Baidu's knowledge-enhanced foundation model takes a giant leap forward. http://research.baidu.com/Blog/index-view?id=185, 2023. Accessed: 2024-05-28.

[Bojar *et al.*, 2018] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[Cettolo *et al.*, 2016] Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Rolando Cattoni, and Marcello Federico. The IWSLT 2016 evaluation campaign. In Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Rolando Cattoni, and Marcello Federico, editors, *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C, December 8-9 2016. International Workshop on Spoken Language Translation.

[Cheng *et al.*, 2018] Minhao Cheng, Jinfeng Yi, Pin Yu Chen, Huan Zhang, and Cho Jui Hsieh. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4), 2018.

[Cheng *et al.*, 2019] Yong Cheng, Lu Jiang, and Wolfgang Macherey. Robust neural machine translation with doubly adversarial inputs. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy, July 2019. Association for Computational Linguistics.

[Cheng *et al.*, 2020] Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. AdvAug: Robust adversarial augmentation for neural machine translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5961–5970, Online, July 2020. Association for Computational Linguistics.

[Cui *et al.*, 2019] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*, 2019.

[Ebrahimi *et al.*, 2018a] Javid Ebrahimi, Daniel Lowd, and Dejing Dou. On adversarial examples for character-level neural machine translation. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[Ebrahimi *et al.*, 2018b] 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)*, pages 31–36, 2018.

[Feng *et al.*, 2022] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, 2022.

[Gao *et al.*, 2023] Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12103–12119, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Johnson *et al.*, 2019] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[Junczys-Dowmunt *et al.*, 2018] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Lai *et al.*, 2022] Siyu Lai, Zhen Yang, Fandong Meng, Xue Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. Generating authentic adversarial examples beyond meaning-preserving with doubly round-trip translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4256–4266, 2022.

[Nakazawa *et al.*, 2016] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios

Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA).

[Ng *et al.*, 2019] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook FAIR's WMT19 news translation task submission. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy, August 2019. Association for Computational Linguistics.

[OpenAI, 2022] OpenAI. ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2022. Accessed: 2024-05-28.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[Sadrizadeh *et al.*, 2023] Sahar Sadrizadeh, AmirHossein Dabiri Aghdam, Ljiljana Dolamic, and Pascal Frossard. Targeted adversarial attacks against neural machine translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Salesky *et al.*, 2023] Elizabeth Salesky, Neha Verma, Philipp Koehn, and Matt Post. Multilingual pixel representations for translation and effective cross-lingual transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13845–13861, 2023.

[Su *et al.*, 2022] Hui Su, Weiwei Shi, Xiaoyu Shen, Zhou Xiao, Tuo Ji, Jiarui Fang, and Jie Zhou. Rocbert: Robust chinese bert with multimodal contrastive pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–931, 2022.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2020] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

[Wang *et al.*, 2021a] Jiakai Wang, Aishan Liu, Xiao Bai, and Xianglong Liu. Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. *IEEE Transactions on Image Processing*, 31:598–611, 2021.

[Wang *et al.*, 2021b] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8565–8574, 2021.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zhang *et al.*, 2021] Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. Crafting adversarial examples for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1967–1977, 2021.