# Enhance Modality Robustness in Text-Centric Multimodal Alignment with Adversarial Prompting

**Yun-Da Tsai**[1], **Ting-Yu Yen**[1], **Keng-Te Liao, Shou-De Lin**

National Taiwan University

f08946007@csie.ntu.edu.tw, r11922042@ntu.edu.tw, d05922001@ntu.edu.tw, sdlin@csie.ntu.edu.tw

**Abstract**

Converting different modalities into generalized text, which then serves as input prompts for large language models (LLMs), is a common approach for aligning multimodal models, particularly when pairwise data is limited. Text-centric alignment method leverages the unique properties of text as a modality space, transforming diverse inputs into a unified textual representation, thereby enabling downstream models to effectively interpret various modal inputs. This study evaluates the quality and robustness of multimodal representations in the face of noise imperfections, dynamic input order permutations, and missing modalities, revealing that current text-centric alignment methods can compromise downstream robustness. To address this issue, we propose a new text-centric adversarial training approach that significantly enhances robustness compared to traditional robust training methods and pre-trained multimodal foundation models. Our findings underscore the potential of this approach to improve the robustness and adaptability of multimodal representations, offering a promising solution for dynamic and real-world applications.

## 1 Introduction

Text-centric multimodal alignment methods have emerged as a powerful approach for integrating multimodal information by converting diverse data types into text. This technique leverages the unique properties of text as a universal modality space, enabling large language models (LLMs) to process and understand visual, auditory, and other forms of data, and have shown competetive performance compared to other traditional embedding-based alignment methods (Tsai et al. 2024). By transforming non-textual information into textual descriptions, these methods facilitate the alignment and integration of various modalities, enhancing the capability of LLMs to comprehend and generate contextually rich responses. For example, LLaVA (Liu et al. 2023c) uses expert models to generate captions, object detection locations, and textual descriptions from images. These are then used as input to GPT-4 to create vision-text instruction-following data as a substitute of actual collecting vision-text instruction-following data, which is inherently difficult and resource-intensive to obtain.



Recent study (Wang et al. 2023b) discovered that Vision LLMs trained on pure synthetically generated high-quality captions by image caption models to replace original noisy data fall into model collapse (Robinson et al. 2021). This phenomenon can be explained by captioning collapse (Vinyals et al. 2015; Wang, Zhang, and Yu 2020) and the one-to-many problem (Young et al. 2014) in image captioning. That is, when transforming images into text, it generates fixed or similar captions for different images, which limits diversity in the output and could potentially jeopardize the downstream model training. This could cause the learned multimodal representations to be less robust for discriminative models (ex. classifier) and cause modality collapse issue for generative models (ex. MLLMs). This leads to the concern where text-centric alignment method would be lead to less robust performance.

In this paper, we improve the modality robustness in text-centric modality alignment methods. Specifically, we aim to repair the modal collapse issue when transforming various modalities into text leads to the generation of fixed or similar outputs, resulting in information loss and reduced diversity. This, in turn, compromises the robustness of the learned multimodal representation. We further propose using adversarial prompting (Yang et al. 2024; Dong et al. 2023; Xu and Wang 2024) and formulate a text-centric adversarial training approach to enhance the modality robustness of text-centric multimodal alignment. Before converting different input modalities into text using expert models and align these modalities within a similar semantic space, we applied a LLM-based perturbation module on top and increase the diversity and robustness of text representations. This adversarial training procedure along with multimodal alignment will optimizes for a more robust performance. This can be easily understand as using LLMs as an adversary to force improve the robustness of multimodal alignment and the downstream model.

In our experiment, different input modalities are converted into text descriptions using expert foundation models for each modality. To evaluate the robustness of these representations, we follow the MULTIBENCH (Liang et al. 2021) framework, which introduces varying levels of and imperfections. This approach simulates real-world conditions, allowing us to assess how well our unified textual representations perform under scenarios of missing or noisy data. By

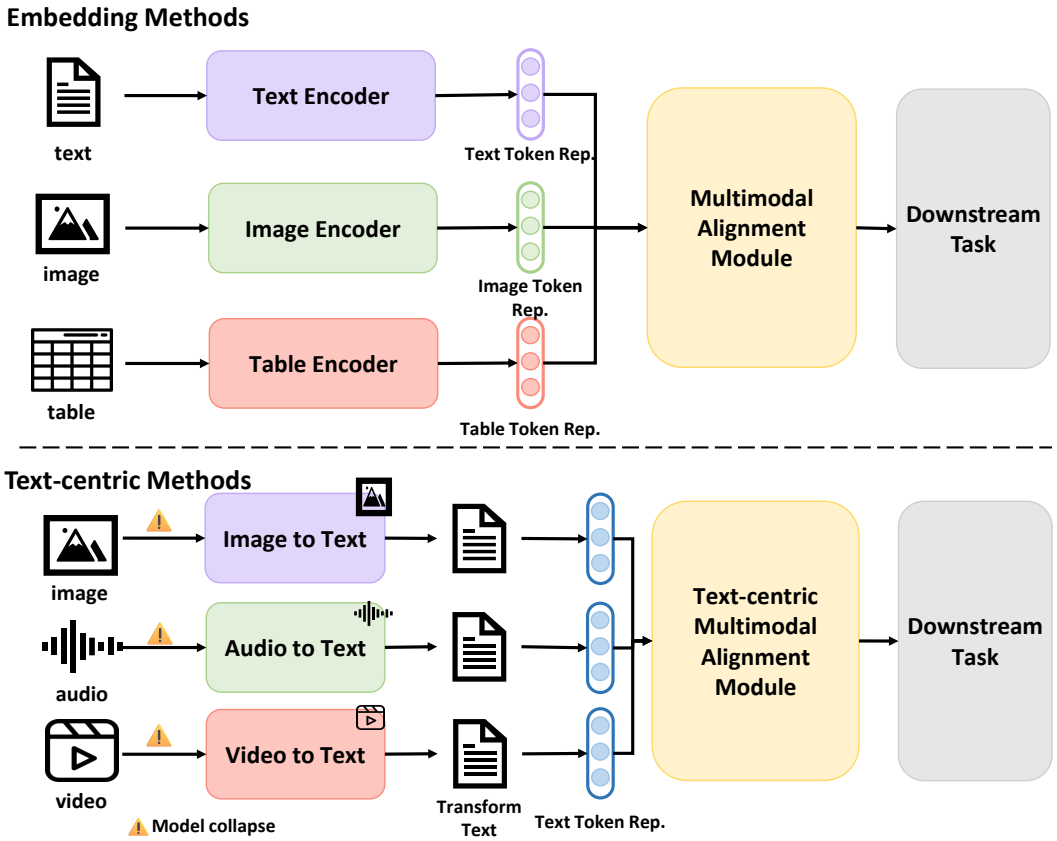


Figure 1: Text-centric multimodal alignment, which converts different modalities into text to serve as input prompts for LLMs, is a common method for aligning large multimodal language models when pairwise multimodal data is limited. The potential model collapse phenomenon can jeopardize the robustness of the aligned representation.

rigorously testing under these conditions, we demonstrate that our enhancement can significantly improve the modality robustness. Qualitative analysis also shows that modality summarization and reasoning augmentation with LLMs offer significant advantages: 1) recovering dropped or corrupted information, 2) transforming implicit relationships into explicit text descriptions, and 3) compensating missing information using LLMs as external knowledge sources. These enhancements contribute to the overall robustness and utility of the multimodal representations.

Our contributions are summarized as follows:

- We are the first to investigate modality robustness in text-centric alignment methods, revealing their inherent lack of robustness.
- We propose an text-centric adversarial training to enhance the robustness for text-centric alignment that demonstrates effective enhancement to modality robustness, consistently outperforming the baselines including traditional robust training methods and multimodal foundation models.
- We provide a qualitative analysis illustrating how large language models (LLMs) strengthen the robustness of textual representations in multimodal alignment.

## 2 Related Work

### 2.1 Text-centric Multimodal Alignment

In recent advancements, several studies have demonstrated the effectiveness of text-centric alignment. For instance, LLaVA (Liu et al. 2023c) utilizes GPT-4 to generate captions and textual descriptions from images, while VideoChat-Text (Li et al. 2023) encodes video content into textual formats. In the medical domain, models like OphGLM (Gao et al. 2023) and ChatCAD (Wang et al. 2023a) extract information from medical images and convert it into diagnostic reports, seamlessly integrating visual data with textual inputs for LLMs. TAMML (Tsai et al. 2024) converts different input modalities into text for downstream model training and demonstrates significant improvements in handling unseen and diverse modality at test time. These approach depends on the quality of transformed text but offers a straightforward way to achieve multimodal integration.

### 2.2 Robustness in Multimodal Learning

Modality robustness (Ma et al. 2022) addresses the issue of different modalities displaying various noise typologies and the potential for real-world multimodal signals to suffer from missing or noisy data in at least one of the modalities.

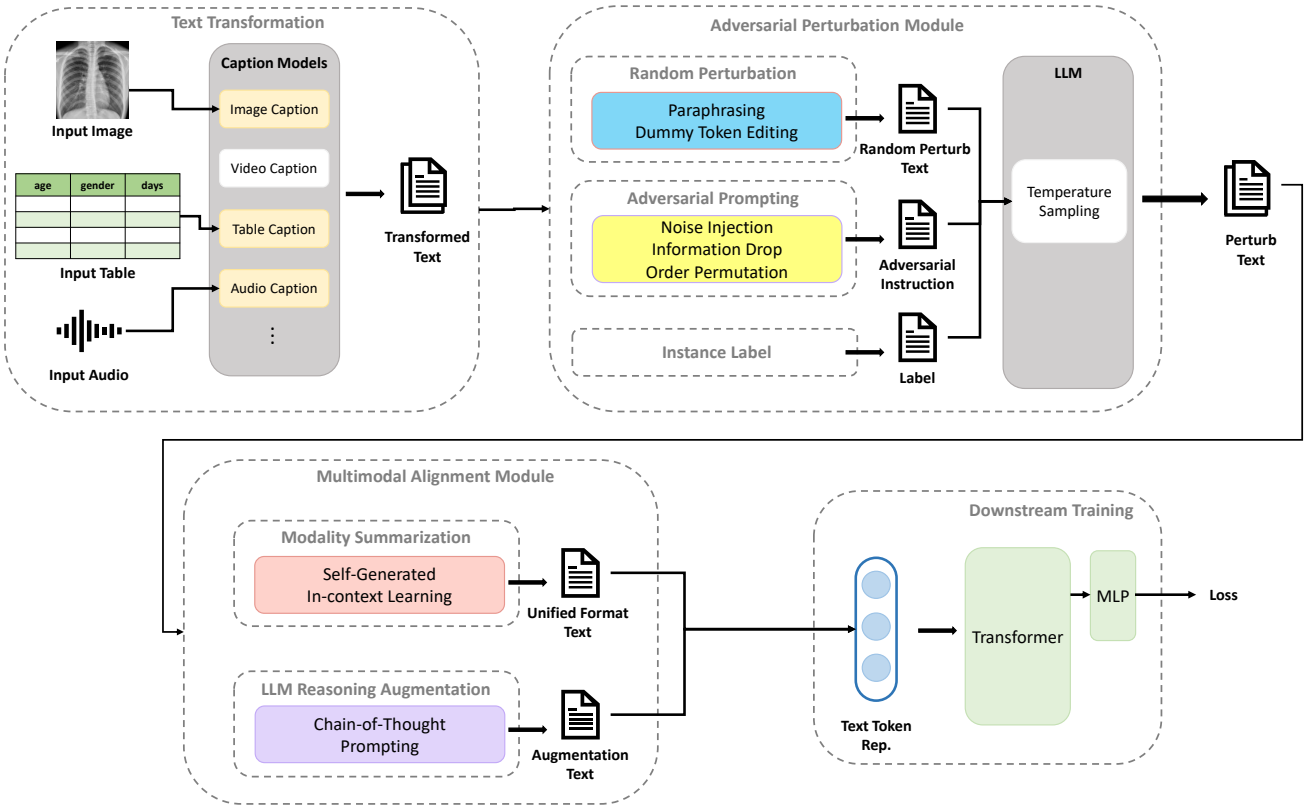


Figure 2: Each raw input modality is transformed into text representations using a corresponding foundation model. Following modality summarization and LLM reasoning are applied in parallel. Finally, the output texts are concatenated as the input to a transformer model for downstream prediction. The inference phase follows a similar pattern. We apply a one-shot in-context learning approach to adapt the linguistic style as anticipated during training.

Similar challenges have been identified in text-centric multimodal alignment methods. Wang et al. (Wang et al. 2023b) discovered that Vision LLMs trained on purely synthetically generated high-quality captions by image caption models, intended to replace original noisy data, suffer from model collapse (Robinson et al. 2021). This phenomenon can be attributed to captioning collapse (Vinyals et al. 2015; Wang, Zhang, and Yu 2020) and the one-to-many problem (Young et al. 2014) in image captioning. When transforming images into text, these models generate fixed or similar captions for different images, limiting diversity in the output and leading to trivial solutions.

### 2.3 Adversarial Prompting

Adversarial prompting exposes vulnerabilities in large language models (LLMs) by manipulating their outputs through various techniques. One such technique, *prompt injection* (Liu et al. 2023b), involves embedding malicious instructions within prompts to alter the intended response of the model, potentially generating harmful or inappropriate content. Another significant method is *prompt leaking* (Perez and Ribeiro 2022; Hui et al. 2024), where crafted prompts extract sensitive information embedded within the model's responses, compromising confidentiality. *Jailbreaking* (Ma et al. 2024; Chao et al. 2023; Liu et al. 2023a) techniques bypass the safety mechanisms of LLMs, enabling the model to produce outputs that violate its ethical guidelines.

Additionally, adversarial prompting has been employed to generate adversarial examples. Techniques such as the Prompt-based Attack Approach (PAT) (Yang et al. 2024; Dong et al. 2023; Xu and Wang 2024) generate adversarial examples via mask-and-filling, exploiting the robustness defects of LLMs. These methods have demonstrated high attack success rates, producing diverse, fluent, and natural adversarial examples that can used to significantly improve the robustness of NLP models.

## 3 Robust Text-centric Multimodal Alignment

This section discusses how we convert raw inputs from different modalities (e.g., images, tabular data) into text representations and apply adversarial prompting to improve the model's robustness. Section 3.1 introduce a text-centic multimodal alignment module. It convert each modality's input into a text representation and align each input modalities. Section 3.2 introduce a perturbation module designed for improving modality robustness by adversarial prompting. The entire process is illustrated in Figure 2.

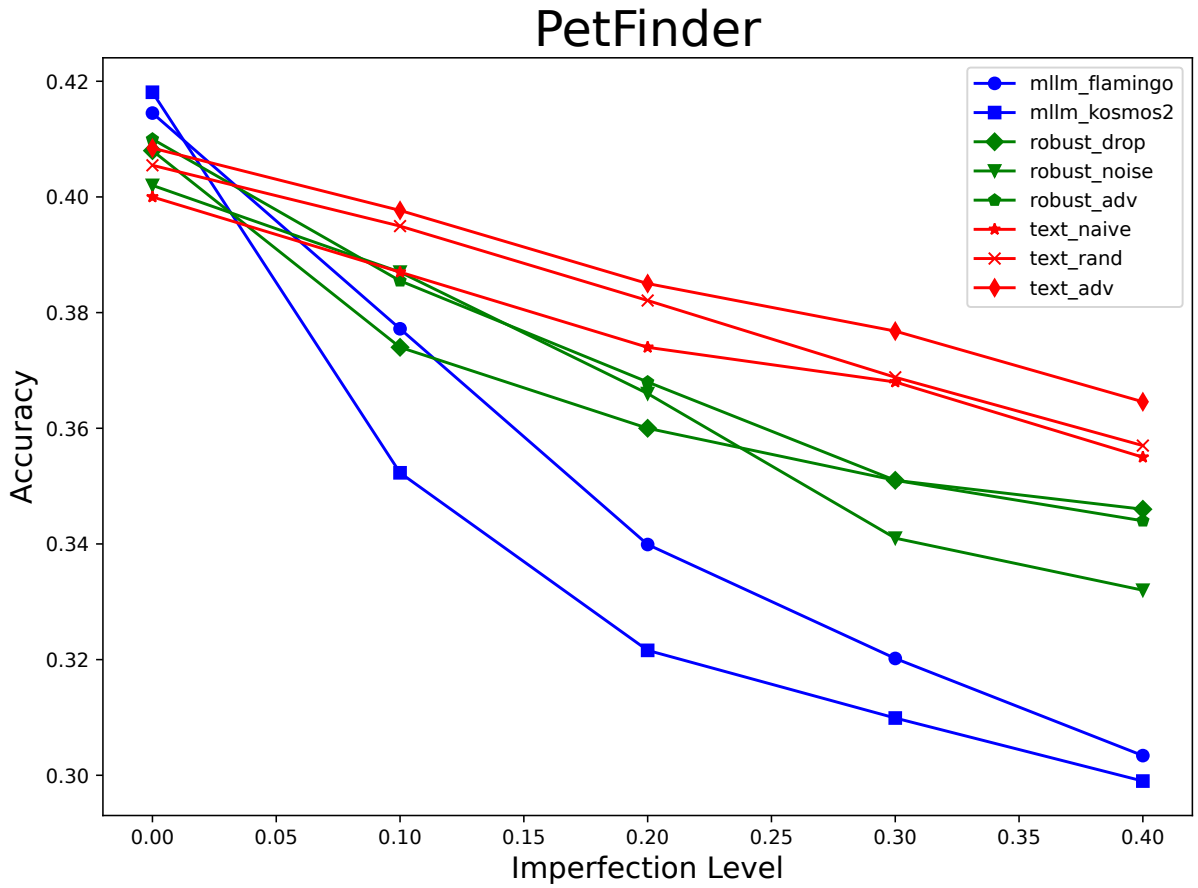


Figure 3: To evaluate the robustness of our model under noisy conditions, we evaluated both relative robustness (top) and effective robustness (bottom) for three datasets. The results from these metrics consistently demonstrate that the text-centric method exhibits superior robustness and resilience to noise when compared to other baseline methods, particularly as noise levels increase. The evaluation was conducted using three different metrics: accuracy, MSE, and RMSE, tailored to each respective dataset.

### 3.1 Text-centic Multimodal Alignment Module

The multimodal alignment module employs LLMs for data transformation across various modalities, aiming to create a unified semantic space. This process is conducted exclusively through in-context learning. Initially, we transform different modalities into text using specialized expert models. Following this, we conduct modality summarization and engage LLMs in text-style translation across modalities. This ensures that all textual representations adopt a consistent linguistic structure, reducing the gap between different modalities and aligning them within a closer semantic space. This step also removes redundant information and mitigates the heterogeneity inherent in text data from diverse sources. Lastly, we include a reasoning augmentation step akin to the Chain-of-Thought method (Wei et al. 2022), enhancing the data with LLMs to boost prediction and judgment capabilities. Moreover, we leverage LLMs as a source of large-scale external knowledge, enriching data understanding and interpretative depth (Chen et al. 2023).

**Text Transformation** In this study, we introduce unique text-based representations for various modalities, enhancing model robustness through rule-based methods and a pretrained MLLM model. We convert raw inputs into a standardized text format, minimizing the need for modality-specific adaptations and reducing data complexity. This method captures vital information while filtering out noise, boosting the model's ability to handle diverse modalities.

For image data, we use a SOTA image captioning model to produce detailed textual descriptions, converting visual content into text. Textual data remains in its original form to preserve linguistic integrity. For tabular data, we apply a simple serialization method from TabLLM (Hegselmann et al. 2023), structured as "The column name is values," proven to surpass zero-shot learning in LLMs. The transformed texts from each modality are then merged and used as input for further processing.

**Modality Summarization** Although all types of data are converted into textual representation, there are still syntactic and semantic gaps between the transformed text across different modalities. In this step, we extend similar lin-

guistic styles text representations to all modalities , improving information quality by facilitating interactions that generate new insights, emphasize key shared data, and remove redundancies. The summary of these modalities is produced by LLMs. Our methodology involves two phases: initially, we collect samples using predefined linguistic styles in prompts that guide the LLMs to merge information from various modalities into a concise summary. Subsequently, this output is integrated with our original prompts, forming a demonstration for in-context learning applied to subsequent samples that follow to the style established in the initial phase.

**LLM Reasoning and Augmentation** We employ LLMs for reasoning based on the Chain-of-Thought method (Wei et al. 2022) and make LLMs as a large-scale external knowledge source similar to (Chen et al. 2023) for data augmentation. By assigning prediction tasks with clear instructions and examples, LLMs analyze and augment the textual inputs based on its external knowledge. The models generate predictions and detailed explanations for each input, enhancing the data through this predictive reasoning process.

### 3.2 Text-centric Perturbation Module

Perturbation module aims at generating more natural, diverse and fluent adversarial examples in order to overcome the model collapse issue and increase robustness. In this section, we introduce two types of perturbation module *Random Perturbation* and *Adversarial Perturbation* that both operates on text using LLMs. Random Perturbation is a naive rule-based prompt perturbation as baseline that randomly derives perturbed variants with Paraphrasing and dummy tokens. Adversarial perturbation employs a set of instruction prompt to guide LLMs in generating adversarial prompts that create the most disruptive examples for various robustness scenarios, effectively shifting the semantics toward the opposite label direction. Both are combined with temperature sampling.

**Random Perturbation** To generate varied inputs, we employ a paraphrasing technique that queries a language model to produce $k$ paraphrased versions of a given input $x_0$. The process is initiated by using the following prompt: *"Suggest $k$ ways to paraphrase the text above. Remain same semantic*. This approach efficiently generates $k$ distinct paraphrased inputs $\{x_i\}$, for $i = 1, \ldots, k$, from a single LLM call, allowing for diverse representations of the original input to be explored. In addition to paraphrasing, we introduce randomness by edit operations including deletion, insertion and substitution with *dummy tokens* to the original input $x_0$. These tokens, denoted by $d_i$, are selected such that they minimally influence the semantic content of the input. Examples of such tokens include newline characters, tab spaces, ellipses, or additional punctuation marks like extra question marks. The modified inputs are represented as $x_i = x_0 + d_i$ or $x_i = d_i + x_0$, where the random perturbations aim to test the model's sensitivity to minor, non-semantic noise.

**Adversarial Perturbation** We designed our adversarial perturbation to simulate various scenarios that challenge the modality robustness of text-centric alignment models. Moreover, the generated examples will intentionally manipulate the input modalities and mislead the model's prediction, thereby creating the most challenging test cases. The process involves crafting specific instruction prompts that guide the language model (LLM) to introduce different types of perturbations into the input data. These perturbations can include noise injection, information dropout, and order permutation, each of which is intended to disrupt the model's understanding and push its predictions toward incorrect labels.

Our approach begins with the pre-generation of a diverse set of instruction prompts specifically tailored for adversarial purposes. These prompts are crafted to induce the LLM to generate adversarial examples that effectively simulate challenging and unpredictable scenarios. The adversarial perturbations are applied modality-wise, allowing us to evaluate and enhance the robustness of the model across different modalities. To systematically generate adversarial examples, we follow these steps:

1. **Random Initialization**: We begin by applying random perturbation to the original input $x$ to create an initial variation $x'$. This step ensures that the base input is already altered before applying further adversarial instructions, increasing the likelihood of generating a significantly misleading example.
2. **Instruction Selection and Parameterization**: An instruction prompt $inst$ is selected from our pre-generated set, which may direct the LLM to perform tasks such as adding noise, dropping critical information, or permuting the order of elements within the input. Alongside the instruction, we set the temperature parameter $T$ to control the randomness of the LLM's output.
3. **Adversarial Example Generation**: The LLM generates the adversarial example $x_{adv}$ by completing the instructed operation in a way that most strongly shifts the semantic content towards the opposite label direction. The result is an adversarial example that challenges the model's ability to maintain accuracy under perturbations. express the process as a formula: $x_{\text{adv}} = \text{LLM}(x', \text{inst}, \text{label}, T)$

This method systematically create samples that simulate real-world scenarios where input modality may be corrupted or misleading. Additionally, this approach supports the iterative refinement of adversarial perturbation.

## 4 Experiment

We conduct experiments on three multimodal dataset and compared baselines including MLLMs, robust training technique and text-centric approaches. We evaluate the robustness under three different scenarios. In all experiments, we use Mixtral 8x7B as default language model and GPT-4-Vision for image captioning, unless specified otherwise. For additional results involving different language models, please refer to Table 3. Furthermore, all trials are run three times and the average is reported. Adversarial training will have maximum ten times more training iterations than regular training.

### 4.1 Dataset

All dataset we used includes three modalities: text, image and tabular. *PetFinder.my Adoption Prediction* (**?**) examines what factors predict how quickly a pet is adopted after being listed. *Airbnb Pricing Prediction* (Cox, Morris, and Higgins 2023) is composed of the following modalities used for making a regression prediction of housing prices. *Avito Demand Prediction* (**?**) predicts the likelihood of an ad selling something based on user item and context features.

### 4.2 Baselines

- **MLLMs:** We selected two state-of-the-art (SOTA) open-source Multimodal Language Models (MLLMs) for robustness comparison: Kosmos-2 (Peng et al. 2023) and Flamingo (Alayrac et al. 2022). This help show the comparison between large foundation models without robust training.
- **Robust Training:** To evaluate the robustness of our text-centric approach against traditional methods, we employed several robust training techniques for the downstream models. These included gaussian noise injection, dropout, and adversarial training using Projected Gradient Descent (PGD) (Madry et al. 2017). These baselines help demonstrate whether our text-based method, which leverages LLMs, offers superior robustness compared to traditional embedding-based methods.
- **Text-Centric Approaches:** We compared the effects of naive(transform to text with no perturbation), random perturbation and adversarial perturbation to determine whether adversarial prompting provides greater robustness than merely increasing input diversity and text transformation.

### 4.3 Evaluation

**Evaluation Protocol** To evaluate the robustness of our models, we adopted similar methodologies outlined in MULTIBENCH (Liang et al. 2021). We define the following three scenarios:

1. Noisy Modality: For images, we introduced Gaussian noise at five different levels from 10% to 90%. For text descriptions, we randomly dropped words with five levels of probability from 10% to 50%. For table data, we randomly dropped column features with probabilities from 10% to 90%.
2. Dynamic Modality: Dynamically permute the order of input modalities to test robustness. Text-centric alignment and token-based transformer models should exhibit invariance to the order of tokens within a prompt.
3. Missing Modality: Randomly select modalities that would be absent at test time. Zero vectors are filled in for robust training.

**Evaluation Metric** Following our evaluation protocols designed to mimic the modality-specific and multimodal imperfections described in MULTIBENCH, we evaluate both *Accuracy*, *MSE*, *RMSE* under imperfections (relative robustness) and the *Drop ratio* of performance when imperfections are introduced (effective robustness).

| | | Petfinder | | Airbnb | | Avito | |
|---|---|---|---|---|---|---|---|
| | | ACC↑ | Drop | MSE↓ | Drop | RMSE↓ | Drop |
| MLLM | Kosmos2 | .371 | .883 | .285 | .954 | .043 | .953 |
| | Flamingo | .374 | .890 | .283 | .961 | .044 | .931 |
| Robust Training | Noise | .296 | .704 | .478 | .569 | .080 | .512 |
| | Dropout | .313 | .745 | .430 | .632 | .067 | .611 |
| | Adv | .302 | .719 | .470 | .578 | .069 | .594 |
| Text centric | Naive | .386 | .919 | .277 | .981 | .042 | .976 |
| | Random | .390 | .928 | .280 | .871 | .043 | .953 |
| | Adv | **.397** | **.945** | **.274** | **.992** | **.042** | **.977** |

Table 1: Dynamic Modality Evaluation. Both relative robustness (left) and effective robustness (right) for three datasets are shown. Text-centric adversarial prompting methods outperforms all baselines and show strong invariance to dynamic input order. Robust training technique completely failed as expected.

| | | Petfinder | | Airbnb | | Avito | |
|---|---|---|---|---|---|---|---|
| | | ACC↑ | Drop | MSE↓ | Drop | RMSE↓ | Drop |
| MLLM | Kosmos2 | .302 | .719 | .320 | .851 | .050 | .824 |
| | Flamingo | .310 | .738 | .318 | .855 | .051 | .803 |
| Robust Training | Noise | .323 | .769 | .319 | .852 | .049 | .836 |
| | Dropout | .310 | .738 | .320 | .850 | .050 | .824 |
| | Adv | .330 | .785 | .308 | .883 | .048 | .854 |
| Text centric | Naive | .362 | .861 | .309 | .880 | .047 | .872 |
| | Random | .370 | .881 | .310 | .877 | .047 | .871 |
| | Adv | **.378** | **.900** | **.302** | **.899** | **.046** | **.891** |

Table 2: Missing Modality Evaluation. Both relative robustness (left) and effective robustness (right) for three datasets are shown. Text-centric adversarial prompting methods outperforms all baselines with a large margin.

### 4.4 Noisy Modality Results

Figure 3 illustrates that our method consistently achieves the lowest drop ratio under noisy modality conditions, outperforming other baselines, particularly at the highest noise levels. For the Petfinder dataset, our text-centric adversarial method experienced only an 11.3% drop, significantly outperforming the robust training method at 15.2% and the MLLMs, which saw a substantial drop of 28.5%. Similar patterns are observed in the Airbnb and Avito datasets, where our method consistently surpasses all baselines. This finding opens a future research direction to explore the text-centric modality collapse problem.

### 4.5 Dynamic Modality Results

To evaluate the model's invariance and robustness to different modality input orders, we tested and averaged the results across all possible input permutations. Table 1 shows that our method has the lowest drop ratio, outperforming all other baselines. For the Petfinder dataset, our text-centric adversarial method experienced only a 5.5% drop in perfor-

| Model | Noisy | Dynamic | Missing |
|---|---|---|---|
| GPT-4o | **0.4086** | **0.398** | **0.381** |
| GPT-3.5-turbo | 0.4037 | **0.398** | 0.379 |
| Mixtral8x7b | 0.4033 | 0.397 | 0.378 |
| w/o alignment | 0.3727 | 0.383 | 0.363 |
| w/o perturbation | 0.3659 | 0.386 | 0.362 |
| w/o both | 0.3342 | 0.373 | 0.302 |

Table 3: Ablation study on each module contribution and the impact of different LLMs on PetFinder dataset. Both alignment module and perturbation module is necessary to perform well. GPT-4o offers the best performance, but the impact between LLMs is not substantial and, at max, ∼2%.

mance, compared to 11.1% for the MLLMs, and far better than robust training methods, which performed close to random guessing, as expected. These trends are consistent in the Airbnb and Avito datasets, where our method consistently outperforms all baselines. The token-based, text-centric approach naturally provides an advantage in maintaining robustness against dynamic input orders, underscoring its effectiveness in various scenarios.

### 4.6 Missing Modality Results

To evaluate robustness under conditions of missing modalities at test-time, we tested and averaged the results across different combinations of dropped modalities. Table 2 shows that our method achieves the lowest drop ratio, outperforming all other baselines. For the Petfinder dataset, our text-centric adversarial method experienced only a 10% drop, which is significantly better than the robust training method at 22.5% and the MLLMs at 26.2%. Similar observations are made in the Airbnb and Avito datasets, where our method consistently outperforms all baselines, reaffirming its superior robustness in scenarios with missing modalities.

### 4.7 Ablation Study

**Module Ablation** We examine the two primary components of our method: the Alignment Module and the Permutation Module. The results presented in Table 3 indicate that both modules contribute almost equally to the overall performance. When both modules are removed, the performance drops significantly, nearing the levels observed with standard MLLMs.

**Language Model Ablation** We adopted different LLMs to test robustness across various LLMs and show that the impact of model type and size is minor, with a maximum difference of around 2% in accuracy. Table 3 shows that GPT-4o offers the best performance among all LLMs.

## 5 Qualitative Analysis and Findings

In this section, we delve into the qualitative aspects and explore how LLMs can effectively handle scenarios where information is either incomplete or lost across different modalities and how LLMs compensate these lost information.

**LLMs recover lost data from other modalities** Figure 4 illustrates that even when critical information is lost from one modality, our approach effectively leverages data from other available modalities to reconstruct the missing content and complementary information across different modalities compensates for gaps, ensuring robust data recovery.

**Text with noise**
The type of pet is Dog. The name of pet is Filo. The age of pet when listed (in months) is 78.The color 1 of pet is Brown. The fur length of pet is Short. The primary breed of pet is Mixed Breed. The secondary breed of pet (if pet is of mixed breed) is Unknown.
**Summarization**
This pet is a gentle and handsome 6-year-old mixed breed dog named Filo. He has a light brown, short coat and soft brown eyes.

Figure 4: The tabular data has dropped the color and fur length column (gray). However, it was recovered (blue) after applying alignment module with LLM that compensate the information from input image.

**Text with noise**
Jack experienced it Jack dog even is He recovering but special care skin which hair to Vet this was to Hopefully lover who take good dog a
**Summarization**
Although the profile write-up for Jack Jack is somewhat fragmented, it indicates that he has experienced some challenges but is on the road to recovery
**Reasoning**
3. **Health Condition**: Jack Jack has a minor injury and is recovering. Potential adopters might be hesitant to take on a pet that requires special care, even if the injury is minor.

Figure 5: This example demonstrates that alignment module can compensate for noisy text input, transforming it into clear descriptions.

**LLMs compensate missing information with knowledge and transform implicit relations into explicit text description by reasoning.** In scenarios where input text is fragmented due to word dropout, and no relevant data is available from other modalities, Figure 5 demonstrates how our method utilizes the extensive knowledge embedded in LLMs.

## 6 Conclusion

This study evaluates the robustness of text-centric multimodal alignment, revealing that it is less robust compared to other robust training techniques. To address this, we propose a novel adversarial training approach specifically designed for text-centric alignment, which outperforms existing baselines that demonstrates strong resistance to noise, input permutation, and missing modalities. Ablation studies further highlight both multimodal alignment and adversarial permutation modules are crucial for enhancing robustness. Additionally, our method is highly transferable across different LLMs. These insights contribute to the development of more resilient multimodal alignment techniques.

## References


Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. 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.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Wang, H.; Zhang, Y.; and Yu, X. 2020. An overview of image caption generation methods. *Computational intelligence and neuroscience*, 2020(1): 3062706.

Robinson, J.; Sun, L.; Yu, K.; Batmanghelich, K.; Jegelka, S.; and Sra, S. 2021. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34: 4974–4986.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M. A.; Zhu, Y.; et al. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*.

Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; and Peng, X. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18177–18186.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Perez, F.; and Ribeiro, I. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.

Chen, H.; Li, Y.; Hong, Y.; Xu, Z.; Gu, Z.; Lan, J.; Zhu, H.; and Wang, W. 2023. Boosting Audio-visual Zero-shot Learning with Large Language Models. *arXiv preprint arXiv: 2311.12268*.

Wang, S.; Zhao, Z.; Ouyang, X.; Wang, Q.; and Shen, D. 2023a. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*.

Cox, M.; Morris, J.; and Higgins, T. 2023. Inside Airbnb: Hawaii. http://insideairbnb.com/get-the-data. Accessed: 2023-09-10.

Chao, P.; Robey, A.; Dobriban, E.; Hassani, H.; Pappas, G. J.; and Wong, E. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; Wang, K.; and Liu, Y. 2023a. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824*.

Gao, W.; Deng, Z.; Niu, Z.; Rong, F.; Chen, C.; Gong, Z.; Zhang, W.; Xiao, D.; Li, F.; Cao, Z.; et al. 2023. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174*.

Liu, Y.; Deng, G.; Li, Y.; Wang, K.; Wang, Z.; Wang, X.; Zhang, T.; Liu, Y.; Wang, H.; Zheng, Y.; et al. 2023b. Prompt Injection attack against LLM-integrated Applications. *arXiv preprint arXiv:2306.05499*.

Dong, X.; He, Y.; Zhu, Z.; and Caverlee, J. 2023. Promptattack: Probing dialogue state trackers with adversarial prompts. *arXiv preprint arXiv:2306.04535*.

Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, 5549–5581. PMLR.

Wang, A. J.; Lin, K. Q.; Zhang, D. J.; Lei, S. W.; and Shou, M. Z. 2023b. Too large; data reduction for vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3147–3157.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023c. Visual Instruction Tuning.

Ma, J.; Cao, A.; Xiao, Z.; Zhang, J.; Ye, C.; and Zhao, J. 2024. Jailbreaking Prompt Attack: A Controllable Adversarial Attack against Diffusion Models. *arXiv preprint arXiv:2404.02928*.

Xu, Y.; and Wang, W. 2024. LinkPrompt: Natural and Universal Adversarial Attacks on Prompt-based Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6473–6486.

Hui, B.; Yuan, H.; Gong, N.; Burlina, P.; and Cao, Y. 2024. PLeak: Prompt Leaking Attacks against Large Language Model Applications. *arXiv preprint arXiv:2405.06823*.

Yang, Y.; Huang, P.; Cao, J.; Li, J.; Lin, Y.; and Ma, F. 2024. A prompt-based approach to adversarial example generation and robustness enhancement. *Frontiers of Computer Science*, 18(4): 184318.

Tsai, Y.-D.; Yen, T.-Y.; Guo, P.-F.; Li, Z.-Y.; and Lin, S.-D. 2024. Text-centric Alignment for Multi-Modality Learning. *arXiv preprint arXiv:2402.08086*.