Supplementary Material of Words or Vision: Do Vision-Language Models Have Blind Faith in Text?

A. Details of Theoretical Analysis

To provide a rigorous foundation for our theoretical analysis, we begin by formally outlining the training process of a vision-language model. For clarity and conciseness, the following is a streamlined adaptation of the standard training process. A VLM is a function $f_{\text{vlm}} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \coloneqq \mathbb{R}^{\tau \times d}$ denotes the set of sequences of d-dimensional feature vector (that can represent text or image) with length τ, and $\mathcal{Y}$ denotes the output space of the model. Without loss of generalization, we assume $\mathcal{Y} \coloneqq \mathbb{R}$ for simplicity.

A.1. Structure

Following Edelman et al. [2], we consider the form of transformer structure of f_{vlm} with L layers as follows. The parameters of i's layer is denoted by $W^{(i)} := \left\{W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}, W_C^{(i)}\right\}$. In addition, we denote $W^{1:i} = \left(W^{(1)}, \ldots, W^{i-1}\right)$ to be the parameters up to i's layer. Further, we let the block of i-th layer $g_{\text{tf-block}}^{(i)} : \mathbb{R}^{\tau \times d} \to \mathbb{R}^{\tau \times d}$ to be

$$
\begin{aligned}
g_{\text{tf-block}}^{(i+1)}\left(X; W^{1:i+1}\right) &:= \Pi_{\text{norm}}\left(\sigma\left(\Pi_{\text{norm}}\left(f\left(X\right)\right)\right) W_C^{(i)}\right) \text{ for } i = 1, \\
g_{\text{tf-block}}^{(i+1)}\left(X; W^{1:i+1}\right) &:= \Pi_{\text{norm}}\left(\sigma\left(\Pi_{\text{norm}}\left(f\left(g_{\text{tf-block}}^{(i)}\left(X; W^{1:i}\right); W^{(i)}\right)\right)\right) W_C^{(i)}\right) \text{ for } i > 1,
\end{aligned}
$$

where $X \in \mathbb{R}^{\tau \times d}$ is the model's input, and Π_{norm} is the layer normalization function, σ is a non-linear activation function, and

$$
f\left(Z; \{W_Q, W_K, W_V\}\right) := \text{ Softmax }\left(Z W_Q \left(Z W_K\right)^\top\right) Z W_V
$$

with $\mathrm{Softmax}(\cdot)$ being the standard softmax function. Finally, the scalar output is defined as

$$
f_{\text{vlm}}(X; W^{1:L}, w) \coloneqq w^\top [g_{\text{tf-block}}^{(L+1)}\left(X; W^{1:L}\right)]_\tau, \text{ for some } w \in \mathbb{R}^d, \tag{1}
$$

where $[\mathbf{G}]_\tau \in \mathbb{R}^d$ denotes the τ-th row of the matrix $\mathbf{G} \in \mathbb{R}^{\tau \times d}$. Furthermore, we have the following assumptions within the structure.

Assumption A.1. For all $i = 1, \cdots, L$, we have $\left\|W_V^{(i)}\right\|_2, \left\|W_K^{(i)} W_Q^{(i)^\top}\right\|_2, \left\|W_C^{(i)}\right\|_2 \leq C_2$.

Assumption A.2. For all $i = 1, \cdots, L$, we have $\left\|W_V^{(i)}\right\|_{2,1}, \left\|W_K^{(i)^\top} W_Q^{(i)}\right\|_{2,1}, \left\|W_C^{(i)}\right\|_{2,1} \leq C_{2,1}$.

Assumption A.3. The activation function $\sigma(\cdot)$ is L_σ-Lipschitz in the l_2 norm.

Assumption A.4. The loss function $l(\cdot)$ is b-bounded and is L_{loss}-Lipschitz in its arguments.

A.2. Training process

Let $\mathcal{X}^{\text{txt}} = [(X_1^{\text{txt}}, y_1^{\text{txt}}), \cdots, (X_N^{\text{txt}}, y_N^{\text{txt}})]$ be a pure-text training set with size N, where $X_i^{\text{txt}} \in \mathbb{R}^{\tau \times d}$ is a sequence of the text feature vector of length τ, and $y_i^{\text{txt}} = f_{\text{gt}}^{\text{txt}}(X_i^{\text{txt}}) \in \mathbb{R}$ is its ground-truth label with $f_{\text{gt}}^{\text{txt}}(\cdot)$ denoted as the ground-true function for the pure text data. We assume $X_1^{\text{txt}}, \cdots, X_N^{\text{txt}}$ are i.i.d. sampled from a unknown distribution $\mathcal{D}^{\text{txt}}$.

In addition, let $\mathcal{X}^{\text{mul}} = [(X_1^{\text{mul}}, y_1^{\text{mul}}), \cdots, (X_N^{\text{mul}}, y_M^{\text{mul}})]$ be a multi-modal training set with size M, where $X_i^{\text{multi}} \in \mathbb{R}^{\tau \times d}$ is a sequence of multi-modal (e.g., text and image) feature vector of length τ, and $y_i^{\text{mul}} = f_{\text{gt}}^{\text{mul}}(X_i^{\text{multi}}) \in \mathbb{R}$ is

its ground-truth label with $f_{\text{gt}}^{\text{mul}}(\cdot)$ denoted as the ground-true function for the multi-modal data. Similarly, we assume $X_1^{\text{mul}}, \cdots, X_N^{\text{mul}}$ are i.i.d. sampled from a unknown distribution $\mathcal{D}^{\text{mul}}$.

Furthermore, let $l : \mathbb{R} \times \mathbb{R} \to$ be a loss function. Then, we define the parameter $\hat{\theta}_{\text{ERM}} \in \Theta$ according to the ERM learning process of the multi-modal paradigm as

$$\hat{\theta}_{\text{ERM}} \in \arg\min_{\theta \in \Theta} \frac{1}{N+M} \left(\sum_{i=1}^{N} l\left(f_{\text{vlm}}(X_i^{\text{txt}}; \theta), y_i^{\text{txt}}\right) + \sum_{i=1}^{M} l\left(f_{\text{vlm}}(X_i^{\text{mul}}; \theta), y_i^{\text{mul}}\right) \right) \tag{2}$$

Our main theoretical result is given in the next subsection.

A.3. Results

We now provide the formal statement of Theorem A.5.

Theorem A.5. *Let Θ be the set of parameters that satisfies Assumption A.1, A.2, A.3 and A.4. For any $\theta \in \Theta$, let $f_{\text{vlm}}(\cdot; \theta)$ be a VLM as is defined in equation 1 with L layers. With probability at least $1 - \delta$,*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\text{txt}}} \left[l\left(f_{\text{vlm}}(X; \hat{\theta}_{\text{ERM}}), f_{\text{gt}}^{\text{txt}}(X)\right) \right] \\
&\lesssim \underbrace{\inf_{\theta \in \Theta} \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\text{txt}}} \left[l\left(f_{\text{vlm}}(X; \theta, f_{\text{gt}}^{\text{txt}}(X)\right) \right]}_{\textit{approximation error}} + \underbrace{\frac{M}{M+N} \sup_{\theta \in \Theta} \left| \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\text{mul}}} \left[l\left(f_{\text{vlm}}(X; \theta), f_{\text{gt}}^{\text{mul}}(X)\right) \right] - \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\text{txt}}} \left[l\left(f_{\text{vlm}}(X; \theta), f_{\text{gt}}^{\text{txt}}(X)\right) \right] \right|}_{\textit{cross-modal error}} \\
&\quad + \underbrace{b\sqrt{\frac{1/\log(\delta)}{N+M}} + L_{\text{loss}} \cdot \sqrt{\frac{C_{\text{vlm}}}{N+M}} \cdot \log(1 + \frac{N+M}{C_{\text{vlm}}})}_{\textit{generalization error}},
\end{aligned} \tag{3}$$

and

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\text{mul}}} \left[l\left(f_{\text{vlm}}(X; \hat{\theta}_{\text{ERM}}), f_{\text{gt}}^{\text{mul}}(X)\right) \right] \\
&\lesssim \underbrace{\inf_{\theta \in \Theta} \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\text{mul}}} \left[l\left(f_{\text{vlm}}(X; \theta, f_{\text{gt}}^{\text{mul}}(X)\right) \right]}_{\textit{approximation error}} + \underbrace{\frac{N}{M+N} \sup_{\theta \in \Theta} \left| \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\text{mul}}} \left[l\left(f_{\text{vlm}}(X; \theta), f_{\text{gt}}^{\text{mul}}(X)\right) \right] - \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\text{txt}}} \left[l\left(f_{\text{vlm}}(X; \theta), f_{\text{gt}}^{\text{txt}}(X)\right) \right] \right|}_{\textit{cross-modal error}} \\
&\quad + \underbrace{b\sqrt{\frac{1/\log(\delta)}{N+M}} + L_{\text{loss}} \cdot \sqrt{\frac{C_{\text{vlm}}}{N+M}} \cdot \log(1 + \frac{N+M}{C_{\text{vlm}}})}_{\textit{generalization error}},
\end{aligned} \tag{4}$$

where

$$C_{\text{vlm}} \lesssim (C_2 L_\sigma)^{O(L)} \cdot B_X^2 B_w^2 C_{2,1}^2 \cdot \log(d\tau(N+M))$$

is the constant related to the covering number of the function class of $\{f_{\text{vlm}}(\cdot; \theta) \mid \theta \in \Theta\}$, and the notation $\lesssim$ hides global constants and logarithmic factors on quantities besides N, M and τ.

A.4. Proof of Theorem A.5

Before we formally prove Theorem A.5, we first present some useful Lemmas from previous works. For any real-valued function class $\mathcal{F}$, we let $\mathcal{N}_\infty\left(\mathcal{F}; \varepsilon; x^{(1)}, \ldots, x^{(m)}\right)$ denote the converting number of $\mathcal{F}$ with respect to the radius ε and the samples $\{x^{(1)}, \ldots, x^{(m)}\}$.

Lemma A.6. *(Adapted from Bartlett and Mendelson [1, Theorem 8] and Edelman et al. [2, Lemma A.4]) Consider a real-valued function class $\mathcal{F}$ such that $|f| \leq A$ for all $f \in \mathcal{F}$ and $\log \mathcal{N}_\infty\left(\mathcal{F}; \varepsilon; x^{(1)}, \ldots, x^{(m)}\right) \leq C_{\mathcal{F}}/\varepsilon^2$ for all $x^{(1)}, \ldots, x^{(m)} \in \mathcal{X}$. Let $l(\cdot, \cdot)$ to be a loss function bounded by b and is L_{loss}-Lipschitz in its arguments, and $g_{\text{gt}} : \mathcal{X} \to \mathbb{R}$ be a ground-true function. Then for any $\delta > 0$ and any distribution $\mathcal{D}$ for the i.i.d samples $x^{(1)}, \ldots, x^{(m)} \in \mathcal{X}$, with probability at least $1 - \delta$, simultaneously for all $f \in \mathcal{F}$,*

$$\left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}} [l(f(x), g_{\text{gt}}(x))] - \frac{1}{m} \sum_{i=1}^{m} l\left(f(x^{(i)}), g_{\text{gt}}(x^{(i)})\right) \right| \leq 4cL_{\text{loss}} \sqrt{\frac{C_{\mathcal{F}}}{m}} \left(1 + \log\left(A\sqrt{m/C_{\mathcal{F}}}\right)\right) + 2b\sqrt{\frac{\log(1/\delta)}{2m}}$$

for some constant $c > 0$.

Lemma A.7. *(Adapted from Edelman et al. [2, Theorem A.17]) Suppose* $\forall i \in [m], \left\|X^{(i)}\right\|_{2,\infty} \leq B_X$. *Let* Θ *be the set of parameters that satisfies Assumption A.1, A.2, A.3 and A.4. For any* $\theta \in \Theta$, *let* $f_{\mathrm{vlm}}(\cdot;\theta)$ *is a vlm model as is fined in equation 1 with* L *layers. We have*

$$\log \mathcal{N}_\infty \left(\{ f_{\mathrm{vlm}}(\cdot;\theta) \mid \theta \in \Theta \}; \varepsilon; X^{(1)}, \ldots, X^{(m)} \right) \lesssim (C_2 L_\sigma)^{O(L)} \cdot \frac{B_X^2 B_w^2 C_{2,1}^2}{\varepsilon^2} \cdot \log(dmT).$$

Proof of Theorem A.5. By Lemma A.6, with probability at least $1-\delta$ we have simultaneously for all $\theta \in \Theta$,

$$\begin{aligned} &\left| \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\mathrm{txt}}} \left[l\left(f_{\mathrm{vlm}}(X;\theta), f_{\mathrm{gt}}^{\mathrm{txt}}(X) \right) \right] - \frac{1}{N} \sum_{i=1}^{N} l\left(f_{\mathrm{vlm}}(X_i^{\mathrm{txt}};\theta), f_{\mathrm{gt}}^{\mathrm{txt}}(X_i^{\mathrm{txt}}) \right) \right| \\ &\leq 4cL_{\mathrm{loss}} \sqrt{\frac{C}{N}} \left(1 + \log\left(A\sqrt{N/C} \right) \right) + 2b \sqrt{\frac{\log(1/\delta)}{2N}}, \end{aligned} \tag{5}$$

where $A \leq (C_2 L_\sigma)^{2L} \cdot B_X$, and C is a constant such that for all $\varepsilon > 0$ and $X^{(1)}, \ldots, X^{(m)} \in \mathbb{R}^{\tau \times d}$ with $\|X^{(i)}\|_{2,\infty} \leq B_X$

$$\log \mathcal{N}_\infty \left(\{ f_{\mathrm{vlm}}(\cdot;\theta) \mid \theta \in \Theta \}; \varepsilon; X^{(1)}, \ldots, X^{(m)} \right) \leq \frac{C}{\varepsilon^2}.$$

Similarly, with probability at least $1-\delta$ we have simultaneously for all $\theta \in \Theta$,

$$\begin{aligned} &\left| \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\mathrm{mul}}} \left[l\left(f_{\mathrm{vlm}}(X;\theta), f_{\mathrm{gt}}^{\mathrm{mul}}(X) \right) \right] - \frac{1}{M} \sum_{i=1}^{M} l\left(f_{\mathrm{vlm}}(X_i^{\mathrm{mul}};\theta), f_{\mathrm{gt}}^{\mathrm{mul}}(X_i^{\mathrm{mul}}) \right) \right| \\ &\leq 4cL_{\mathrm{loss}} \sqrt{\frac{C}{M}} \left(1 + \log\left(A\sqrt{M/C} \right) \right) + 2b \sqrt{\frac{\log(1/\delta)}{2M}}. \end{aligned} \tag{6}$$

Note that for any $\theta \in \Theta$ we have

$$\begin{aligned} &\left| \frac{1}{N+M} \left(\sum_{i=1}^{N} l\left(f_{\mathrm{vlm}}(X_i^{\mathrm{txt}};\theta), y_i^{\mathrm{txt}} \right) + \sum_{i=1}^{M} l\left(f_{\mathrm{vlm}}(X_i^{\mathrm{mul}};\theta), y_i^{\mathrm{mul}} \right) \right) - \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\mathrm{txt}}} \left[l\left(f_{\mathrm{vlm}}(X;\theta), f_{\mathrm{gt}}^{\mathrm{txt}}(X) \right) \right] \right| \\ &= \left| \frac{N}{M+M} \left(\frac{1}{N} \sum_{i=1}^{N} l\left(f_{\mathrm{vlm}}(X_i^{\mathrm{txt}};\theta), y_i^{\mathrm{txt}} \right) - \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\mathrm{txt}}} \left[l\left(f_{\mathrm{vlm}}(X;\theta), f_{\mathrm{gt}}^{\mathrm{txt}}(X) \right) \right] \right) \right. \\ &\quad \left. + \frac{M}{M+N} \left(\frac{1}{M} \sum_{i=1}^{M} l\left(f_{\mathrm{vlm}}(X_i^{\mathrm{mul}};\theta), y_i^{\mathrm{mul}} \right) - \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\mathrm{txt}}} \left[l\left(f_{\mathrm{vlm}}(X;\theta), f_{\mathrm{gt}}^{\mathrm{txt}}(X) \right) \right] \right) \right| \end{aligned} \tag{7}$$

$$\begin{aligned} &\overset{(a)}{\leq} \frac{N}{M+M} \left| \frac{1}{N} \sum_{i=1}^{N} l\left(f_{\mathrm{vlm}}(X_i^{\mathrm{txt}};\theta), y_i^{\mathrm{txt}} \right) - \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\mathrm{txt}}} \left[l\left(f_{\mathrm{vlm}}(X;\theta), f_{\mathrm{gt}}^{\mathrm{txt}}(X) \right) \right] \right| \\ &\quad + \frac{M}{M+N} \left| \frac{1}{M} \sum_{i=1}^{M} l\left(f_{\mathrm{vlm}}(X_i^{\mathrm{mul}};\theta), y_i^{\mathrm{mul}} \right) - \mathop{\mathbb{E}}_{X \sim \mathcal{D}^{\mathrm{txt}}} \left[l\left(f_{\mathrm{vlm}}(X;\theta), f_{\mathrm{gt}}^{\mathrm{txt}}(X) \right) \right] \right|, \end{aligned} \tag{8}$$

where (a) follows from Jensen's inequality.

In addition, with probability at least $1-\delta$,we have for all $\theta \in \Theta$

$$\begin{aligned}
&\left|\frac{1}{M}\sum_{i=1}^{M} l\left(f_{\text{vlm}}(X_i^{\text{mul}};\theta), y_i^{\text{mul}}\right) - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{txt}}(X)\right)\right]\right| \\
&\leq \left|\frac{1}{M}\sum_{i=1}^{M} l\left(f_{\text{vlm}}(X_i^{\text{mul}};\theta), y_i^{\text{mul}}\right) - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{mul}}(X)\right)\right]\right| \\
&\quad + \left|\mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{txt}}(X)\right)\right] - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{mul}}(X)\right)\right]\right| \\
&\overset{(a)}{\leq} 4cL_{\text{loss}}\sqrt{\frac{C}{M}}\left(1+\log\left(A\sqrt{M/C}\right)\right) + 2b\sqrt{\frac{\log(1/\delta)}{2M}} \\
&\quad + \left|\mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{txt}}(X)\right)\right] - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{mul}}(X)\right)\right]\right|
\end{aligned} \tag{9}$$

where (a) follows from equation 6.

Combining equation 5, equation 8 and equation 9, we get that with probability at least $1-\delta$, for all $\theta \in \Theta$,

$$\begin{aligned}
&\left|\frac{1}{N+M}\left(\sum_{i=1}^{N} l\left(f_{\text{vlm}}(X_i^{\text{txt}};\theta), y_i^{\text{txt}}\right) + \sum_{i=1}^{M} l\left(f_{\text{vlm}}(X_i^{\text{mul}};\theta), y_i^{\text{mul}}\right)\right) - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{txt}}(X)\right)\right]\right| \\
\leq & 4cL_{\text{loss}}\frac{\sqrt{MC}}{N+M}\left(1+\log\left(A\sqrt{M/C}\right)\right) + 2b\frac{\sqrt{M\log(1/\delta)/2}}{N+M} \\
& + 4cL_{\text{loss}}\frac{\sqrt{NC}}{N+M}\left(1+\log\left(A\sqrt{N/C}\right)\right) + 2b\frac{\sqrt{N\log(1/\delta)/2}}{N+M} \\
& + \left|\mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{txt}}(X)\right)\right] - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{mul}}(X)\right)\right]\right|.
\end{aligned} \tag{10}$$

By the definition of $\hat{\theta}_{\text{ERM}}$, equation 10 implies that with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\frac{1}{N+M}\left(\sum_{i=1}^{N} l\left(f_{\text{vlm}}(X_i^{\text{txt}};\hat{\theta}_{\text{ERM}}), y_i^{\text{txt}}\right) + \sum_{i=1}^{M} l\left(f_{\text{vlm}}(X_i^{\text{mul}};\hat{\theta}_{\text{ERM}}), y_i^{\text{mul}}\right)\right) - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\hat{\theta}_{\text{ERM}}), f_{\text{gt}}^{\text{txt}}(X)\right)\right]\right| \\
\leq & \inf_{\theta\in\Theta} \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\theta, f_{\text{gt}}^{\text{mul}}(X)\right)\right] + \frac{N}{M+N}\sup_{\theta\in\Theta}\left|\mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{mul}}(X)\right)\right] - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{txt}}(X)\right)\right]\right| \\
& + 4cL_{\text{loss}}\frac{\sqrt{MC}}{N+M}\left(1+\log\left(A\sqrt{M/C}\right)\right) + 2b\frac{\sqrt{M\log(1/\delta)/2}}{N+M} + 4cL_{\text{loss}}\frac{\sqrt{NC}}{N+M}\left(1+\log\left(A\sqrt{N/C}\right)\right) + 2b\frac{\sqrt{N\log(1/\delta)/2}}{N+M} \\
& + \left|\mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\hat{\theta}_{\text{ERM}}), f_{\text{gt}}^{\text{txt}}(X)\right)\right] - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\hat{\theta}_{\text{ERM}}), f_{\text{gt}}^{\text{mul}}(X)\right)\right]\right|.
\end{aligned} \tag{11}$$

Note the fact that $\max\{N,M\} \leq N+M \leq 2\max\{N,M\}$. Finally, by Lemma A.7 and hiding global constants and logarithmic factors on quantities besides N, M and τ, we get with probability $1-\delta$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\hat{\theta}_{\text{ERM}}), f_{\text{gt}}^{\text{txt}}(X)\right)\right] \\
&\lesssim \inf_{\theta\in\Theta} \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\theta, f_{\text{gt}}^{\text{txt}}(X)\right)\right] + \frac{M}{M+N}\sup_{\theta\in\Theta}\left|\mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{mul}}(X)\right)\right] - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{txt}}(X)\right)\right]\right| \\
&\quad + b\sqrt{\frac{1/\log(\delta)}{N+M}} + L_{\text{loss}}\cdot\sqrt{\frac{C_{\text{vlm}}}{N+M}}\cdot\log(1+\frac{N+M}{C_{\text{vlm}}}),
\end{aligned} \tag{12}$$

and similarly,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\hat{\theta}_{\text{ERM}}), f_{\text{gt}}^{\text{mul}}(X)\right)\right] \\
&\lesssim \inf_{\theta\in\Theta} \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\theta, f_{\text{gt}}^{\text{mul}}(X)\right)\right] + \frac{N}{M+N}\sup_{\theta\in\Theta}\left|\mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{mul}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{mul}}(X)\right)\right] - \mathop{\mathbb{E}}_{X\sim\mathcal{D}^{\text{txt}}}\left[l\left(f_{\text{vlm}}(X;\theta), f_{\text{gt}}^{\text{txt}}(X)\right)\right]\right| \\
&\quad + b\sqrt{\frac{1/\log(\delta)}{N+M}} + L_{\text{loss}}\cdot\sqrt{\frac{C_{\text{vlm}}}{N+M}}\cdot\log(1+\frac{N+M}{C_{\text{vlm}}}).
\end{aligned} \tag{13}$$

This completes the proof of Theorem A.5 □

B. Experimental Setup

This section outlines the experimental setup, including examples of constructed textual variations, details of the brand detection task [4], and the evaluation protocols employed. We present examples illustrating the three types of textual variations alongside the corresponding image, original question, and ground-truth answers to provide clarity and context.

B.1. Examples

This subsection provides examples of matching, corrupted, and irrelevant texts across different datasets in Tables 1 to 4.

	**Q:** What green veggie is on the pizza	**GT:** pepper
Match:	The pizza has green pepper slices on one of its sections.	
Corruption:	The pizza has green broccoli florets on one of its sections.	
Irrelevance:	Beckham obtained his early education at Roseland Academy in Bardstown. In 1881 he served as a page in the Kentucky House of Representatives at the age of 12. Later, he enrolled at Central University (now Eastern Kentucky University) in Richmond, Kentucky but was forced to quit school at the age of 17 to support his widowed mother. Two years later, he became principal of Bardstown public schools, serving from 1888 to 1893. Concurrently, he studied law at the University of Kentucky, where he earned his law degree in 1889. He was admitted to the bar and commenced practice in Bardstown in 1893. He also served as president of the Young Democrats ' Club of Nelson County .	

Table 1. Illustration of matching, corrupted, and irrelevant information in a sample from VQAv2.

B.2. Brand Recognition

Brand recognition from a webpage is a crucial step in detecting phishing websites. Phishing webpages aim to deceive users by imitating the appearance of legitimate websites associated with well-known brands. Accurately identifying the brand linked to a webpage allows for a comparison between the input webpage's URL and the official URL of the recognized brand, aiding in the detection of phishing attempts.

In our experiments, we utilized phishing webpage samples from the TR-OP dataset [4]. Each sample comprises a screenshot and its corresponding HTML code. Depending on the scenario, the HTML content either reflects the target brand displayed in the screenshot or is altered to assess the model's robustness. We evaluated three specific scenarios:

- **Matching:** The original HTML includes information about the target brand visible in the screenshot. This scenario provides the model with consistent inputs, helping it correctly identify the brand.
- **Corruption:** In this case, we inserted a fabricated brand name (e.g., "The official webpage of MobrisPremier") into the HTML to mislead the model into recognizing a non-existent brand. Since no corresponding URL exists for such brands, phishing detection becomes infeasible for these inputs.
- **Irrelevance:** The HTML content was replaced with randomly selected sentences from the Wiki dataset [], ensuring that the new content was unrelated to any brand. This scenario tests the model's ability to handle inputs with no brand-specific information.

To standardize the inputs, we preprocessed the HTML content by removing all tags and truncating it to a maximum length of 5,000 characters.

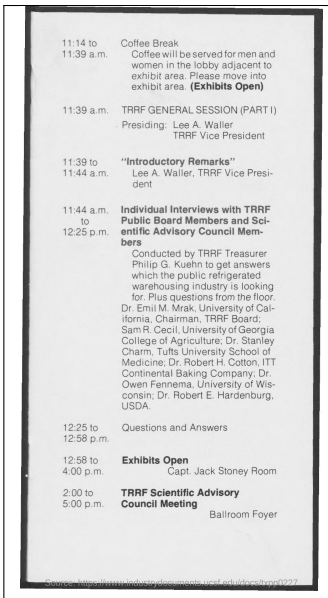	**Q:** What time is ‘question and answers ‘session?	**GT:** 12:25 to 12:58 p.m.
Match:	The ’Questions and Answers’ session is scheduled from 12:25 to 12:58 p.m.	
Corruption:	The ’Questions and Answers’ session is scheduled from 2:00 to 5:00 p.m.	
Irrelevance:	The Americans knew of the approach of the Japanese forces from reports from native scouts and their own patrols , but did not know exactly where or when they would attack . The ridge around which Edson deployed his men consisted of three distinct hillocks . At the southern tip and surrounded on three sides by thick jungle was Hill 80 (so named because it rose 80 ft (24 m) above sea level) . Six hundred yards north was Hill 123 (123 ft (37 m) high) , the dominant feature on the ridge . The northernmost hillock was unnamed and about 60 ft (18 m) high . Edson placed the five companies from the Raider battalion on the west side of the ridge and the three Parachute battalion companies on the east side , holding positions in depth from Hill 80 back to Hill 123 . Two of the five Raider companies , B̈ änd C̈ ,̈ held a line between the ridge , a small , swampy lagoon , and the Lunga River . Machine @-@ gun teams from Ë C̈ompany , the heavy weapons company , were scattered throughout the defenses . Edson placed his command post on Hill 123 .	

Table 2. Illustration of matching, corrupted, and irrelevant information in a sample from DocVQA.

B.3. Evaluation

We follow the evaluation protocol specified for each dataset. To reduce cases where models generate open-ended answers, which complicates evaluation, we adopt a similar approach to the evaluation setting in `LLaVA-1.5` [5]. For certain datasets, we append additional formatting prompts after the question, as shown in Table 5.

For MathVista [6], which uses GPT-based evaluation, we do not include formatting prompts. Instead, GPT is employed directly to evaluate the outputs.

C. Experimental Results

To rigorously assess the performance impact of varying textual contexts, we have documented the comprehensive results across four distinct datasets. These results are quantified using several metrics: Accuracy, Normalized Accuracy, and Text Preference Ratio (TPR) for the text variations of Match, Corruption, and Irrelevance, alongside Macro Accuracy. The detailed outcomes are encapsulated in Table 6.

For a thorough assessment of the investigated methodologies, encompassing base models, instructional prompts, and Supervised Fine-Tuning (SFT), we present results across four datasets, measured in terms of Accuracy, Normalized Accuracy,

	**Q:** Hint: Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end. Question: what is the total volume of the measuring cup? (Unit: g)	**GT:** 1000
Match:	The measuring cup has markings up to 1000 grams, indicating its total volume capacity.	
Corruption:	The measuring cup has markings up to 500 grams, indicating its total volume capacity.	
Irrelevance:	The windmill at Thelnetham was built by millwright George Bloomfield for William Button in 1819 . It replaced a post mill which had been moved to Sandy Lane , Diss , Norfolk the previous year . The mill was set to work on Christmas Day 1819 . In 1832 , the mill was modernised by the fitting of a cast @-@ iron windshaft , Patent sails and a fantail . The new windshaft was fitted on 16 July 1832 . It was made by J Aickman , the Kings Lynn millwright , and weighs 1 0̆0be tons (1 @,@ 780 kg) . A new stock was fitted in September 1836 . William Button died on 11 February 1837 . The mill passed jointly to his widow Rebecca and their son Richard . Richard Button worked the mill until 1860 , at which date it was conveyed to his sons Richard and William , who sold it to Richard Peverett from Kenninghall , Norfolk in 1862.	

Table 3. Illustration of matching, corrupted, and irrelevant information in a sample from MathVista.

Text Preference Ratio (TPR) under the text variations of Match, Corruption, and Irrelevance, as well as Macro Accuracy. These experiments were conducted utilizing the models `LLaVA-NeXT-7B` and `Qwen2-VL-7B`. The detailed findings are provided in Table 7.

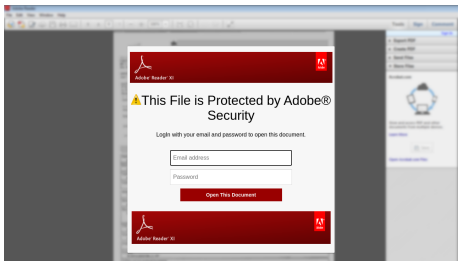	**Q:** Define targeted brand as a brand that a webpage belongs to. Given the screenshot of a webpage P as the primary information for identifying the target brand and the text as additional reference, determine what the targeted brand of P is. The text can be HTML from the webpage or something irrelevant. Please be careful with the text, as it may contain noise or adversarial attacks. You must output the targeted brand of P even if you are not sure about it. Only output the brand name without any additional information. Input information: –HTML: "html" Answer:	**GT:** Adobe
Match:	Adobe® PDF Reader® Xl This File is Protected by Adobe® Security LogIn with your email and password to open this document.	
Corruption:	The official webpage of MobrisPremier. The official webpage of MobrisPremier. Adobe® PDF Reader® Xl This File is Protected by Adobe® Security LogIn with your email and password to open this document.	
Irrelevance:	Ayola began her career in the theatre , performing in a number of Shakespearean plays including Twelfth Night , A Midsummer Night 's Dream , The Tempest , The Merchant of Venice and Macbeth . She states of this : " Shakespeare keeps coming my way . I love the fact that I get to play people who are much more articulate than I 'll ever be " . Ayola has performed in Twelfth Night in the lead roles of both Olivia and Viola . She explains : " The role of Viola didn 't sit that well with me for some reason but Olivia makes more sense . " She has also appeared in modern performances , assuming the title role of Dido , Queen of Carthage at the Globe Theatre in London in 2003 , which she described as " a dream of a part " . She has deemed her dream role to be that of Isabella in Measure for Measure , as she once lost out on the part and would like to prove herself capable of playing it.	

Table 4. Illustration of matching, corrupted, and irrelevant information in a sample from Brand Recognition.

Dataset	**Response Formatting Prompts**
VQAv2 [3]	Please only output the answer with a single word or phrase.
DocVQA [7]	Please only output the answer directly.
MathVista [6]	–
Brand Recognition [4]	Only output the brand name without any additional information.

Table 5. Response formatting prompts used for evaluation.

Model	Base ↑	Match			Corruption			Irrelevance			Macro ↑
		Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR ↓	Accuracy ↑	Norm ↑	TPR ↓	
GPT-4o mini	69.82	87.49	125.31	89.15	51.55	73.83	52.42	72.11	103.28	3.77	70.38
Claude Haiku	51.02	82.81	**162.31**	86.74	26.33	51.61	82.71	51.10	100.16	13.95	53.41
GPT-4o	78.39	89.27	113.88	69.03	**70.75**	90.25	27.09	78.82	100.55	1.56	**79.61**
Claude Sonnet	66.88	77.85	116.40	49.86	68.17	**101.93**	**9.58**	70.89	**106.00**	1.38	72.30
LLaVA-NeXT-7B	79.45	92.32	116.20	86.25	28.69	36.11	85.52	79.43	99.97	4.72	66.81
LLaVA-NeXT-13B	81.02	**93.59**	115.51	86.45	37.61	46.42	74.43	81.29	100.33	3.30	70.83
LLaVA-NeXT-34B	82.96	93.07	112.19	79.10	42.87	51.68	67.56	79.64	95.99	2.70	71.86
Phi3.5	75.65	91.23	120.59	79.51	35.23	46.57	74.05	74.87	98.97	2.25	67.11
Molmo-7B-D	76.33	88.57	116.04	88.32	49.29	64.57	59.40	76.50	100.22	9.36	71.45
Qwen2-VL-7B	**85.51**	92.76	108.48	13.17	50.79	59.40	29.22	**83.70**	97.88	**1.28**	75.75

(a) **VQAv2**

Model	Base ↑	Match			Corruption			Irrelevance			Macro ↑
		Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR ↓	Accuracy ↑	Norm ↑	TPR ↓	
GPT-4o mini	69.40	81.40	117.26	82.74	38.20	55.04	52.07	67.20	96.83	0.80	62.27
Claude Haiku	69.53	83.45	120.06	68.77	39.35	56.61	47.67	57.82	83.16	1.18	60.21
GPT-4o	85.00	90.40	106.35	64.75	73.60	86.59	17.96	86.40	**101.65**	0.23	83.47
Claude Sonnet	87.00	91.53	105.15	41.18	**84.60**	**97.24**	**3.21**	87.41	100.47	**0.00**	**87.85**
LLaVA-NeXT-7B	53.60	90.80	**169.40**	86.92	10.00	18.66	87.77	52.40	97.76	0.71	51.07
LLaVA-NeXT-13B	57.70	90.40	156.68	87.82	11.00	19.06	86.84	55.80	96.68	0.65	52.40
LLaVA-NeXT-34B	64.00	87.80	137.19	84.62	15.10	23.59	82.69	62.70	97.97	0.13	55.20
Phi3.5	78.20	92.40	118.16	58.01	50.50	64.60	40.51	77.00	98.46	**0.00**	73.30
Molmo-7B-D	74.00	90.30	122.30	87.54	38.40	51.89	57.20	74.70	100.95	0.37	67.80
Qwen2-VL-7B	**90.50**	**95.10**	105.08	51.97	57.50	63.64	37.41	**89.90**	99.34	0.22	80.83

(b) **DocVQA**

Model	Base ↑	Match			Corruption			Irrelevance			Macro ↑
		Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR ↓	Accuracy ↑	Norm ↑	TPR ↓	
GPT-4o mini	52.30	73.80	141.11	88.82	23.90	45.70	80.28	44.40	84.89	20.14	47.37
Claude Haiku	41.00	**80.30**	195.85	88.04	19.80	48.29	77.42	39.70	96.83	23.33	46.60
GPT-4o	**58.90**	73.70	125.04	85.20	41.20	69.95	48.98	53.10	90.15	13.55	56.00
Claude Sonnet	56.30	68.10	120.95	57.69	**49.30**	**87.57**	**29.14**	**55.20**	98.05	**7.96**	**57.53**
LLaVA-NeXT-7B	35.80	74.80	**273.62**	88.72	19.70	54.97	84.19	28.40	**104.02**	38.22	40.97
LLaVA-NeXT-13B	36.20	76.20	257.43	88.98	20.60	56.89	80.83	32.60	96.28	37.18	43.13
LLaVA-NeXT-34B	34.00	68.00	200.00	73.59	21.70	61.98	67.64	32.10	94.41	20.40	40.60
Phi3.5	43.10	73.70	171.21	84.82	22.20	51.47	80.20	41.10	95.36	13.99	45.67
Molmo-7B-D	44.90	68.50	152.57	82.46	32.90	73.27	60.63	45.30	100.89	27.49	48.90
Qwen2-VL-7B	55.40	77.80	140.43	84.50	28.90	52.18	70.23	54.90	99.10	8.44	53.87

(c) **MathVista**

Model	Base ↑	Match			Corruption			Irrelevance			Macro ↑
		Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR ↓	Accuracy ↑	Norm ↑	TPR ↓	
GPT-4o mini	88.84	86.88	97.80	30.43	84.80	95.44	7.48	88.48	99.60	0.08	86.72
Claude Haiku	84.40	83.40	98.81	26.02	78.72	93.27	6.44	82.28	97.49	**0.00**	81.47
GPT-4o	88.68	89.48	100.90	14.64	89.76	**101.22**	**0.83**	89.16	**100.54**	0.04	89.47
Claude Sonnet	**90.20**	**90.56**	100.40	17.03	**90.24**	100.04	0.96	**90.24**	100.04	**0.00**	**90.35**
LLaVA-NeXT-7B	78.60	77.56	98.67	82.39	62.52	79.54	64.74	16.28	20.72	70.45	52.12
LLaVA-NeXT-13B	83.00	79.00	95.18	77.04	33.96	40.92	72.97	11.72	14.12	79.61	41.56
LLaVA-NeXT-34B	66.28	68.28	**102.99**	31.60	53.52	80.77	23.49	52.84	79.69	10.65	58.21
Phi3.5	84.40	83.84	99.33	31.39	60.68	71.90	50.54	16.44	19.48	79.17	53.65
Molmo-7B-D	87.44	87.32	99.86	37.38	41.44	47.39	60.40	60.88	69.63	27.36	63.21
Qwen2-VL-7B	89.68	88.92	99.15	17.22	86.48	96.43	2.99	70.16	78.20	15.73	81.85

(d) **Brand Detection**

Table 6. Performance in Accuracy, Normalized Accuracy (Norm) and Text Preference Ratio (TPR) across four datasets under three text variations: Match, Corruption, and Irrelevance. The **Macro** column represents the average of Match, Corruption, and Irrelevance **Accuracy** for each model, calculated to be comparable to the **Base** accuracy.

Model	Base ↑	Match			Corruption			Irrelevance			Macro ↑
		Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR	
LLaVA-NeXT-7B	**79.45**	92.32	116.20	86.25	28.69	36.11	85.52	79.43	99.97	4.72	66.81
Instruction	**79.45**	92.25	116.12	86.46	34.27	43.13	78.50	78.15	98.36	6.69	68.22
SFT	77.48	87.56	113.01	59.73	71.25	91.94	20.00	77.32	99.79	4.06	**78.71**
Qwen2-VL-7B	**85.51**	92.76	108.48	13.17	50.79	59.40	29.22	83.70	97.88	1.28	75.75
Instruction	**85.51**	92.62	108.32	14.42	54.78	64.07	27.01	82.82	96.85	1.18	76.74
SFT	84.18	87.01	103.36	36.65	82.72	98.26	6.69	84.00	99.79	2.59	**84.58**

(a) **VQAv2**

Model	Base ↑	Match			Corruption			Irrelevance			Macro ↑
		Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR	
LLaVA-NeXT-7B	**53.60**	90.80	169.40	86.92	10.00	18.66	87.77	52.40	97.76	0.71	51.07
Instruction	**53.60**	88.60	165.30	84.01	9.80	18.28	87.38	49.40	92.16	1.54	49.27
SFT	52.20	75.50	144.63	56.21	42.80	81.99	28.19	50.20	96.17	0.14	**56.17**
Qwen2-VL-7B	**90.50**	95.10	105.08	51.97	57.50	63.64	37.41	89.90	99.34	0.22	80.83
Instruction	**90.50**	94.70	104.64	51.46	57.80	63.88	37.00	89.80	99.23	0.11	80.77
SFT	90.30	93.10	103.10	26.06	84.30	93.35	6.32	89.50	99.11	0.11	**88.97**

(b) **DocVQA**

Model	Base ↑	Match			Corruption			Irrelevance			Macro ↑
		Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR	
LLaVA-NeXT-7B	**35.80**	74.80	208.94	84.32	19.70	55.03	84.19	28.40	79.33	34.57	41.03
Instruction	**35.80**	70.60	197.77	84.68	21.80	60.89	81.85	31.20	87.15	32.94	41.20
SFT	35.30	68.70	194.90	77.42	23.50	66.57	63.75	32.70	92.64	10.76	**41.63**
Qwen2-VL-7B	55.40	77.80	140.43	84.50	28.90	52.17	70.23	54.90	99.10	8.44	53.87
Instruction	55.40	78.10	140.79	86.50	29.30	52.88	70.59	54.90	99.10	8.11	54.10
SFT	**58.50**	74.00	126.50	78.31	40.30	68.89	49.16	57.20	97.78	5.65	**57.17**

(c) **MathVista**

Model	Base ↑	Match			Corruption			Irrelevance			Macro ↑
		Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR	Accuracy ↑	Norm ↑	TPR	
LLaVA-NeXT-7B	78.60	77.56	98.67	68.30	62.52	79.54	59.17	16.28	20.72	89.14	46.44
Instruction	78.60	78.36	99.70	66.57	54.84	69.77	59.63	8.88	11.30	85.26	47.36
SFT	**81.36**	78.32	96.26	37.18	69.48	85.39	17.92	69.08	84.92	9.08	**72.29**
Qwen2-VL-7B	**89.68**	88.92	99.15	17.22	86.48	96.43	2.99	70.16	78.20	15.73	81.85
Instruction	**89.68**	88.52	98.71	17.50	87.12	97.15	1.94	77.80	86.77	9.34	84.48
SFT	89.44	90.08	100.72	20.32	88.76	99.24	1.43	87.40	97.72	0.71	**88.75**

(d) **Brand Detection**

Table 7. Performance of investigated solutions in Accuracy, Normalized Accuracy (Norm) and Text Preference Ratio (TPR) across four datasets under three text variations: Match, Corruption, and Irrelevance. The **Macro** column represents the average of Match, Corruption, and Irrelevance **Accuracy** for each model, calculated to be comparable to the **Base** accuracy.

References

[1] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002. 2

[2] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022. 1, 2, 3

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 8

[4] Yuexin Li, Chengyu Huang, Shumin Deng, Mei Lin Lock, Tri Cao, Nay Oo, Bryan Hooi, and Hoon Wei Lim. Knowphish: Large language models meet multimodal knowledge graphs for enhancing reference-based phishing detection. *arXiv preprint arXiv:2403.02253*, 2024. 5, 8

[5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 6

[6] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. 6, 8

[7] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 8