A Detailed Proof of the Lemmas and Theorems

A.1 Proof of the Theorem 4.1

Proof. We just following the proof in the Ge et al. (2023). First, we reformulate the optimization problem in equation 4 as

$$\hat{\phi}_{\mathcal{A}} = \underset{\phi \in \Phi_{\mathcal{A}}}{\arg\max} \sum_{i,j=1}^{n_{\mathcal{A}}} \log p_{\phi}(\phi(\boldsymbol{x}_{ij}^{\mathcal{A}}), s_{ij}) \tag{14}$$

where $\phi(\boldsymbol{x}_{ij}) = (z_i^{\mathcal{A}}, z_j^{\mathcal{A}}) = (\phi(x_i^{\mathcal{A}}), \phi(x_j^{\mathcal{A}}))$, and

$$p_{\phi}(\phi(\boldsymbol{x}_{ij}^{\mathcal{A}}), s_{ij}) = \frac{\exp(z_i^{\mathcal{A}} \cdot z_j^{\mathcal{A}} / \tau)}{\sum_t \exp(z_t^{\mathcal{A}} \cdot z_j^{\mathcal{A}} / \tau)} \tag{15}$$

the Gibbs distribution for the paired data. In fact, it just formulates the cross-modality contrastive learning framework by the maximum likelihood estimation (MLE). We ignore the $\mathcal{A}$ subscripts/upscripts and the side information s for notation simplicity in the proof. By the definition of $\hat{\phi}$, we have

$$0 \leq \frac{1}{2}\left(\sum_{i,j=1}^{n} \log p_{\hat{\phi}}(\boldsymbol{x}_{ij}) - \sum_{i,j=1}^{n} \log p_{\phi^*}(\boldsymbol{x}_{ij})\right) \tag{16}$$

$$= \frac{1}{2}\sum_{i,j=1}^{n} \log \frac{p_{\hat{\phi}}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})} \tag{17}$$

$$\tag{18}$$

To construct the relationship between d_{TV} and the previous formula, we use Markov inequality and Boole inequality (subadditivity of events). Recall that we define $\mathcal{P}_{\mathcal{X}\times\mathcal{S}}(\Phi) = \{p_{\phi}(\boldsymbol{x}, s) | \phi \in \Phi\}$ as the possible distribution family of Φ. For notation simplicity, we denote $\mathcal{P}_{\mathcal{X}_{\mathcal{A}}\times\mathcal{S}}(\Phi_{\mathcal{A}})$ as $\mathcal{P}$ in this proof. Then we denote the ϵ-bracket class as $\mathcal{N}_{[]}(\mathcal{P}, \epsilon)$, $N_{[]}(\mathcal{P}, \epsilon) = |\mathcal{N}_{[]}(\mathcal{P}, \epsilon)|$. For any $\overline{p}_{\phi} \in \mathcal{N}_{[]}(\mathcal{P}, \epsilon)$, we have the following Markov inequality,

$$\mathbb{P}(\exp(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\phi}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})} \geq t)) \leq \frac{\mathbb{E}[\exp(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\hat{\phi}}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})})]}{t} \tag{19}$$

$$\mathbb{P}\left(\exp\left(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\phi}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})} \geq \frac{C\mathbb{E}[\exp(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\hat{\phi}}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})})]}{\delta}\right)\right) \leq \delta / C \tag{20}$$

$$\tag{21}$$

Define the event $D_{\overline{p}_{\phi}}$ as

$$D_{\overline{p}_{\phi}} = \{\boldsymbol{x} : \exp\left(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\phi}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})}\right) \geq \frac{C\mathbb{E}[\exp(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\phi}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})})]}{\delta}\} \tag{22}$$

Then by iterating over all $\overline{p}_{\phi} \in \mathcal{N}_{[]}(\mathcal{P}_A, \epsilon)$ we have,

$$\mathbb{P}(\cup_{\overline{p}_{\phi} \in \mathcal{N}_{[]}(\mathcal{P}_A, \epsilon)} D_{\overline{p}_{\phi}}) \leq \sum_{\overline{p}_{\phi} \in \mathcal{N}_{[]}(\mathcal{P}_A, \epsilon)} \mathbb{P}(D_{\overline{p}_{\phi}}) \tag{23}$$

$$\leq \frac{N_{[]}(\mathcal{P}_A, \epsilon) \cdot \delta}{C} \tag{24}$$

Take $C = N_{[]}(\mathcal{P}, \epsilon)$, we have with probability at least $1 - \delta$, for all $\overline{p}_{\phi} \in \mathcal{N}_{[]}(\mathcal{P}, \epsilon)$

$$\exp\left(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\phi}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})}\right) \leq \mathbb{E}[\exp(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\phi}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})})] \cdot \frac{N_{[]}(\mathcal{P}, \epsilon)}{\delta} \tag{25}$$

$$\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\phi}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})} \leq \log \mathbb{E}[\exp(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\phi}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})})] + \log \frac{N_{[]}(\mathcal{P}, \epsilon)}{\delta} \tag{26}$$

By the definition of bracket class, $\overline{p}_{\hat{\phi}}$ satisfies, with probability at least $1-\delta$

$$\begin{aligned}
0 \leq \frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\hat{\phi}}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})} &\leq \log \mathbb{E}[\exp(\frac{1}{2}\sum_{i,j=1}^{n} \log \frac{\overline{p}_{\hat{\phi}}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})})] + \log \frac{N_{[]}(\mathcal{P},\epsilon)}{\delta} && (27)\\
&= \sum_{i,j=1}^{n} \log \mathbb{E}[\sqrt{\frac{\overline{p}_{\hat{\phi}}(\boldsymbol{x}_{ij})}{p_{\phi^*}(\boldsymbol{x}_{ij})}}] + \log \frac{N_{[]}(\mathcal{P},\epsilon)}{\delta} && (28)\\
&= m^2 \log \int \sqrt{\overline{p}_{\hat{\phi}}(\boldsymbol{x}) \cdot p_{\phi^*}(\boldsymbol{x})} d\boldsymbol{x} + \log \frac{N_{[]}(\mathcal{P},\epsilon)}{\delta} && (29)\\
&\leq m^2 (\int \sqrt{\overline{p}_{\hat{\phi}}(\boldsymbol{x}) \cdot p_{\phi^*}(\boldsymbol{x})} d\boldsymbol{x} - 1) + \log \frac{N_{[]}(\mathcal{P},\epsilon)}{\delta} && (30)\\
& && (31)
\end{aligned}$$

By rearranging the terms,

$$\begin{aligned}
1 - \int \sqrt{\overline{p}_{\hat{\phi}}(\boldsymbol{x}) \cdot p_{\phi^*}(\boldsymbol{x})} d\boldsymbol{x} &\leq \frac{1}{m^2} \log \frac{N_{[]}(\mathcal{P}_A,\epsilon)}{\delta} && (32)\\
\int \left(\sqrt{\overline{p}_{\hat{\phi}}(\boldsymbol{x})} - \sqrt{p_{\phi^*}(\boldsymbol{x})}\right)^2 d\boldsymbol{x} &\leq \frac{2}{m^2} \log \frac{N_{[]}(\mathcal{P}_A,\epsilon)}{\delta} && (33)\\
& && (34)
\end{aligned}$$

By the definition of ϵ-bracket class, we have

$$\begin{aligned}
\int \left(\sqrt{\overline{p}_{\hat{\phi}}(\boldsymbol{x})} + \sqrt{p_{\phi^*}(\boldsymbol{x})}\right)^2 dx &\leq 2 + 2\int \sqrt{\overline{p}_{\hat{\phi}}(\boldsymbol{x}) \cdot p_{\phi^*}(\boldsymbol{x})} dx && (35)\\
&\leq 2 + \int \overline{p}_{\hat{\phi}}(\boldsymbol{x}) + p_{\phi^*}(\boldsymbol{x}) dx && (36)\\
&\leq 2 + 2(\epsilon + 1) = 2\epsilon + 4 && (37)
\end{aligned}$$

Now we can bound the d_{TV} by Cauchy-Schwarz inequality, with probability at least $1-\delta$

$$\begin{aligned}
d_{TV}\left(\mathbb{P}_{\hat{\phi}}(\boldsymbol{x}), \mathbb{P}_{\phi^*}(\boldsymbol{x})\right) &= \frac{1}{2}\int |p_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})| d\boldsymbol{x} && (38)\\
&\leq \frac{1}{2}\int |\overline{p}_{\hat{\phi}}(\boldsymbol{x}) - p_{\phi^*}(\boldsymbol{x})| d\boldsymbol{x} + \frac{1}{2}\int |p_{\hat{\phi}}(\boldsymbol{x}) - \overline{p}_{\hat{\phi}}(\boldsymbol{x})| d\boldsymbol{x} && (39)\\
&\leq \frac{1}{2}\left(\int \left(\sqrt{\overline{p}_{\hat{\phi}}(\boldsymbol{x})} - \sqrt{p_{\phi^*}(\boldsymbol{x})}\right)^2 d\boldsymbol{x} \cdot \int \left(\sqrt{\overline{p}_{\hat{\phi}}(\boldsymbol{x})} + \sqrt{p_{\phi^*}(\boldsymbol{x})}\right)^2 d\boldsymbol{x}\right)^{1/2} + \frac{\epsilon}{2} && (40)\\
&\leq \frac{1}{2}\sqrt{\frac{2}{n^2}\log\frac{N_{[]}(\mathcal{P},\epsilon)}{\delta} \cdot (2\epsilon + 4)} + \frac{\epsilon}{2} && (41)
\end{aligned}$$

set $\epsilon = \frac{1}{m^2}$ we can bound the formula above by

$$d_{TV}\left(\mathbb{P}_{\hat{\phi}}(\boldsymbol{x}), \mathbb{P}_{\phi^*}(\boldsymbol{x})\right) \leq 3\sqrt{\frac{1}{n^2}\log\frac{N_{[]}(\mathcal{P},\frac{1}{n^2})}{\delta}} \tag{42}$$

□

A.2 PROOF OF THE THEOREM 4.2

Proof. The proof is mainly from Chap. 6 in Zhang (2023).
First, we define

$$\epsilon(\mathcal{L} \circ \Phi_{\mathcal{B}}, S_m^2) = \sup_{\phi_{\mathcal{B}} \in \Phi_{\mathcal{B}}} [\mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{A}}, \phi_{\mathcal{B}}, \boldsymbol{x}, s)] - \frac{1}{m^2}\sum_{i,j=1}^{m} \mathcal{L}(\hat{\phi}_{\mathcal{A}}, \phi_{\mathcal{B}}, \boldsymbol{x}_{ij}, s_{ij})] \tag{43}$$

and

$$\epsilon_n(\mathcal{L} \circ \Phi_{\mathcal{B}}) = \mathbb{E}_{S_m^2}\, \epsilon(\mathcal{L} \circ \Phi_{\mathcal{B}}, S_m^2) \tag{44}$$

where $S_m^2 = \{(x_i^{\mathcal{A}}, x_i^{\mathcal{B}})\}_m \times \{(x_i^{\mathcal{A}}, x_i^{\mathcal{B}})\}_m$. Then by the symmetrization, we can prove

$$\epsilon_n(\mathcal{L} \circ \Phi_{\mathcal{B}}) \le 2R_n(\mathcal{L} \circ \Phi_{\mathcal{B}}) \tag{45}$$

we do not give detailed proof, readers can refer to Theorem 6.3 in Zhang (2023). Consider $f(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{mm}) = \epsilon(\mathcal{L} \circ \Phi_{\mathcal{B}}, S_m^2)$, by the assumption that $\sup_{\phi_{\mathcal{B}} \in \Phi_{\mathcal{B}}, \boldsymbol{x}_{ij}} \langle \phi_{\mathcal{B}}(\boldsymbol{x}_i), \phi_{\mathcal{B}}(\boldsymbol{x}_j) \rangle \le B$, we can check the condition for McDiarmid's inequality,

$$\sup_{\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{mm}, \boldsymbol{x}'_{ij}} |f(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}'_{ij}, \ldots, \boldsymbol{x}_{mm}) - f(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}'_{ij}, \ldots, \boldsymbol{x}_{mm})| \tag{46}$$

$$\le \frac{1}{m^2} \sup_{\boldsymbol{x}_{ij}, \boldsymbol{x}'_{ij}} |\sup_{\phi_{\mathcal{B}}}[\mathcal{L}(\hat{\phi}_{\mathcal{A}}, \phi_{\mathcal{B}}, \boldsymbol{x}'_{ij}, s'_{ij}) - \mathcal{L}(\hat{\phi}_{\mathcal{A}}, \phi_{\mathcal{B}}, \boldsymbol{x}_{ij}, s_{ij})]| \tag{47}$$

$$\le \frac{1}{m^2} \sup_{\boldsymbol{x}_{ij}, \boldsymbol{x}'_{ij}} \sup_{\phi_{\mathcal{B}}} |\mathcal{L}(\hat{\phi}_{\mathcal{A}}, \phi_{\mathcal{B}}, \boldsymbol{x}'_{ij}, s'_{ij}) - \mathcal{L}(\hat{\phi}_{\mathcal{A}}, \phi_{\mathcal{B}}, \boldsymbol{x}_{ij}, s_{ij})| \tag{48}$$

$$\le \frac{1}{m^2} 2 |\log \frac{1 + \exp(B)}{1 + \exp(-B)}| \tag{49}$$

$$= \frac{2B}{m^2} \tag{50}$$

$$\tag{51}$$

then we can apply McDiarmid's inequality,

$$\mathbb{P}(f(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{mm}) \ge \mathbb{E}_{S_m^2} f(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{mm}) + \epsilon) \le \exp(\frac{-m^2 \epsilon^2}{2B^2}) \tag{52}$$

then with probability at least $1 - \delta$,

$$f(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{mm}) \le \mathbb{E}_{S_m^2} f(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{mm}) + B\sqrt{\frac{2\ln(1/\delta)}{m^2}} \tag{53}$$

$$\sup_{\phi_{\mathcal{B}} \in \Phi_{\mathcal{B}}} [\mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{A}}, \phi_{\mathcal{B}}, \boldsymbol{x}, s)] - \frac{1}{m^2} \sum_{i,j=1}^{m} \mathcal{L}(\hat{\phi}_{\mathcal{A}}, \phi_{\mathcal{B}}, \boldsymbol{x}_{ij}, s_{ij})] \le \epsilon_n(\mathcal{L} \circ \Phi_{\mathcal{B}}) + B\sqrt{\frac{2\ln(1/\delta)}{m^2}} \tag{54}$$

Combined with the result of equation 45, we have with probability at least $1 - \delta$, for any $\phi_{\mathcal{B}} \in \Phi_{\mathcal{B}}$,

$$\mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{A}}, \hat{\phi}_{\mathcal{B}}, \boldsymbol{x}, s)] - \frac{1}{m^2} \sum_{i,j=1}^{m} \mathcal{L}(\hat{\phi}_{\mathcal{A}}, \hat{\phi}_{\mathcal{B}}, \boldsymbol{x}_{ij}, s_{ij}) \le 2R_n(\mathcal{L} \circ \Phi_{\mathcal{B}}) + B\sqrt{\frac{2\ln(1/\delta)}{m^2}} \tag{55}$$

A similar discussion shows that with probability at least $1 - \delta$, $\phi_{\mathcal{B}} \in \Phi_{\mathcal{B}}$,

$$\frac{1}{m^2} \sum_{i,j=1}^{m} \mathcal{L}(\hat{\phi}_{\mathcal{A}}, \hat{\phi}_{\mathcal{B}}, \boldsymbol{x}_{ij}, s_{ij}) - \mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{A}}, \hat{\phi}_{\mathcal{B}}, \boldsymbol{x}, s)] \le 2R_n(\mathcal{L} \circ \Phi_{\mathcal{B}}) + B\sqrt{\frac{2\ln(1/\delta)}{m^2}} \tag{56}$$

□

A.3 PROOF OF THE THEOREM 4.3

We first introduce a lemma.

Lemma A.1 (Bound of ERM.)**.** *Suppose that $\mathcal{L}(\cdot, \cdot)$ is a L-bounded loss function. Given a fixed $\phi \in \Phi$, with probability at least $1 - \delta$, for any $\psi \in \Psi$,*

$$\mathbb{E}[\mathcal{L}(\psi \circ \phi(x), y)] - \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\psi \circ \phi(x), y) \le 2R_n(\mathcal{L} \circ \Psi \circ \phi) + L\sqrt{\frac{2\ln(1/\delta)}{n}} \tag{57}$$

Proof. This proof is almost the same as the proof in Theorem 4.2. First, we define $\epsilon(\mathcal{L}\circ\Psi\circ\phi, S_n) = \sup_{\psi\in\Psi}[\mathbb{E}[\mathcal{L}(\psi\circ\phi(x),y)] - \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\psi\circ\phi(x),y)]$ and $\epsilon_n(\mathcal{L}\circ\Psi\circ\phi) = \mathbb{E}_{S_n}\epsilon(\mathcal{L}\circ\Psi\circ\phi, S_n)$ where $S_n = \{(x_i,y_i)\}_n$. Then by the symmetrization, we can prove

$$\epsilon_n(\mathcal{L}\circ\Psi\circ\phi) \le 2R_n(\mathcal{L}\circ\Psi\circ\phi) \tag{58}$$

we do not give detailed proof, readers can refer to Theorem 6.3 in Zhang (2023). Consider $f(X_1,\dots,X_n) = \sup_{\psi\in\Psi}[\mathbb{E}[\mathcal{L}(\psi\circ\phi(x),y)] - \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\psi\circ\phi(x),y)]$, it is obvious that $\sup_{x_1,\dots,x_n,x_i'}|f(x_1,\dots,x_i,\dots,x_n) - f(x_1,\dots,x_i',\dots,x_n)| \le \frac{2}{n}L$, then we can apply McDiarmid's inequality,

$$\mathbb{P}(f(x_1,\dots,x_n) \ge \mathbb{E}_{S_n}f(x_1,\dots,x_n) + \epsilon) \le \exp(\frac{-n\epsilon^2}{2L^2}) \tag{59}$$

then with probability at least $1-\delta$,

$$f(x_1,\dots,x_n) \le \mathbb{E}_{S_n}f(x_1,\dots,x_n) + L\sqrt{\frac{2\ln(1/\delta)}{n}} \tag{60}$$

$$\sup_{\psi\in\Psi}[\mathbb{E}[\mathcal{L}(\psi\circ\phi(x),y)] - \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\psi\circ\phi(x),y)] \le \epsilon_n(\mathcal{L}\circ\Psi\circ\phi) + L\sqrt{\frac{2\ln(1/\delta)}{n}} \tag{61}$$

Combined with the result of equation 45, we have with probability at least $1-\delta$, for any $\psi\in\Psi$,

$$\mathbb{E}[\mathcal{L}(\psi\circ\phi(x),y)] - \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\psi\circ\phi(x),y) \le 2R_n(\mathcal{L}\circ\Psi\circ\phi) + L\sqrt{\frac{2\ln(1/\delta)}{n}} \tag{62}$$

A similar discussion shows that with probability at least $1-\delta$, for any $\psi\in\Psi$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\psi\circ\phi(x),y) - \mathbb{E}[\mathcal{L}(\psi\circ\phi(x),y)] \le 2R_n(\mathcal{L}\circ\Psi\circ\phi) + L\sqrt{\frac{2\ln(1/\delta)}{n}} \tag{63}$$

take $1-\delta/2$ for each inequality and combine the results, we get with probability at least $1-\delta$, for any $\psi\in\Psi$,

$$|\mathbb{E}[\mathcal{L}(\psi\circ\phi(x),y)] - \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\psi\circ\phi(x),y)| \le 2R_n(\mathcal{L}\circ\Psi\circ\phi) + L\sqrt{\frac{2\ln(2/\delta)}{n}} \tag{64}$$

□

Now we prove the Theorem 4.3,

Proof. The proof starts from the standard convergence analysis with Rademacher complexity. By the Lemma A.1, given the fixed $\hat{\phi}_{\mathcal{B}}(x)$, we have with probability at least $1-\delta$,

$$\mathbb{E}[\mathcal{L}(\hat{\psi}_{\mathcal{B}}\circ\hat{\phi}_{\mathcal{B}}(x),y)] \le \frac{1}{n}\sum_{i=1}^{n_{\mathcal{B}}}\mathcal{L}(\hat{\psi}_{\mathcal{B}}\circ\hat{\phi}_{\mathcal{B}}(x),y) + 2R_n(\mathcal{L}\circ\Psi_{\mathcal{B}}\circ\hat{\phi}_{\mathcal{B}}) + L\sqrt{\frac{2\ln(1/\delta)}{n_{\mathcal{B}}}} \tag{65}$$

we only need to handle the empirical risk $\frac{1}{n}\sum_{i=1}^{n_{\mathcal{B}}}\mathcal{L}(\hat{\phi}_{\mathcal{B}},\hat{\psi}_{\mathcal{B}}(x),y)$, from the definition of $\hat{\psi}_{\mathcal{B}}$ in equation 6 we get,

$$\frac{1}{n}\sum_{i=1}^{n_{\mathcal{B}}}\mathcal{L}(\hat{\psi}_{\mathcal{B}}\circ\hat{\phi}_{\mathcal{B}}(x),y) \le \epsilon_{\mathcal{B}} + \frac{1}{n}\sum_{i=1}^{n_{\mathcal{B}}}\mathcal{L}(\psi^*_{\mathcal{B}}\circ\hat{\phi}_{\mathcal{B}}(x),y) - \mathbb{E}[\mathcal{L}(\psi^*_{\mathcal{B}}\circ\hat{\phi}_{\mathcal{B}}(x),y)] \tag{66}$$

$$+\mathbb{E}[\mathcal{L}(\psi^*_{\mathcal{B}}\circ\hat{\phi}_{\mathcal{B}}(x),y)] \tag{67}$$

the first term equation 66 can be bounded by concentration inequality and we only need to bound the second term equation 67 further. By the assumption 4.1,

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{B}}, \psi^*_{\mathcal{B}}(x), y)] &\le \kappa \mathbb{E}[\mathbb{E}_{x'}[\mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{B}}, (x, x'), s)]] \tag{68}\\
&= \kappa \mathbb{E}[\mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{B}}, (x, x'), s)] \quad (\text{denote } \boldsymbol{x} = (x, x')) \tag{69}\\
&= \kappa(\mathbb{E}[\mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{B}}, \boldsymbol{x}, s)] - \mathbb{E}[\mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{A}}, \boldsymbol{x}, s)] \tag{70}\\
&\quad + \mathbb{E}[\mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{A}}, \boldsymbol{x}, s)] - \mathbb{E}[\mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)] \tag{71}\\
&\quad + \mathbb{E}[\mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)]) \tag{72}\\
&= \kappa(\mathbb{E}[-p_{\phi^*_{\mathcal{B}}}(\boldsymbol{x}, s) \log p_{\hat{\phi}_{\mathcal{B}}}(\boldsymbol{x}, s) + p_{\phi^*_{\mathcal{A}}}(\boldsymbol{x}, s) \log p_{\hat{\phi}_{\mathcal{B}}}(\boldsymbol{x}, s)] \tag{73}\\
&\quad + \mathbb{E}[-p_{\phi^*_{\mathcal{A}}}(\boldsymbol{x}, s) \log p_{\hat{\phi}_{\mathcal{B}}}(\boldsymbol{x}, s) + p_{\hat{\phi}_{\mathcal{A}}}(\boldsymbol{x}, s) \log p_{\hat{\phi}_{\mathcal{B}}}(\boldsymbol{x}, s)] \tag{74}\\
&\quad + \mathbb{E}[\mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)]) \tag{75}\\
&\le \kappa B \cdot \Big(d_{TV}(\mathbb{P}_{\phi^*_{\mathcal{B}}}, \mathbb{P}_{\phi^*_{\mathcal{A}}}) + d_{TV}(\mathbb{P}_{\hat{\phi}_{\mathcal{A}}}, \mathbb{P}_{\phi^*_{\mathcal{A}}}) \Big) \tag{76}\\
&\quad + \kappa \mathbb{E}[\mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)] \tag{77}\\
& \tag{78}
\end{aligned}$$

where the inequality equation 104 comes from a same argument as equation 49.

To derive the final result, define two events,

$$D_{\mathcal{A}} = \left\{ S^{\mathcal{A}}_n : d_{TV}(\mathbb{P}_{\hat{\phi}_{\mathcal{A}}}(\boldsymbol{x}, s), \mathbb{P}_{\phi^*_{\mathcal{A}}}(\boldsymbol{x}, s)) \le 3\sqrt{\frac{1}{{n_{\mathcal{A}}}^2} \log \frac{N_{[]}(\mathcal{P}_{\mathcal{X}_{\mathcal{A}} \times \mathcal{S}}(\Phi_{\mathcal{A}}), \frac{1}{{n_{\mathcal{A}}}^2})}{\delta}} \right\} \tag{79}$$

$$D_{\mathcal{A}\mathcal{B}} = \left\{ S^2_m : \mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)] - \frac{1}{m^2} \sum_{i,j=1}^{m} \mathcal{L}_{\text{CMD}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}_{ij}, s_{ij}) \le 2R_n(\mathcal{L}_{\text{CMD}} \circ \Phi_{\mathcal{B}}) + L\sqrt{\frac{2\ln(1/\delta)}{m^2}} \right\} \tag{80}$$

By the Theorem 4.2 and Lemma 4.1, we have $\mathbb{P}(D_{\mathcal{A}}) \ge 1 - \delta, \mathbb{P}(D_{\mathcal{A}\mathcal{B}} | \hat{\phi}_{\mathcal{A}}) \ge 1 - \delta$, then consider the $\mathbb{P}(D_{\mathcal{A}} \cap D_{\mathcal{A}\mathcal{B}})$,

$$\begin{aligned}
\mathbb{P}(D_{\mathcal{A}} \cap D_{\mathcal{A}\mathcal{B}}) &= \mathbb{E}[\mathbb{1}_{D_{\mathcal{A}}} \mathbb{P}(D_{\mathcal{A}\mathcal{B}} | \hat{\phi}_{\mathcal{A}})] \tag{81}\\
&\ge (1 - \delta) \cdot \mathbb{P}(D_{\mathcal{A}}) \tag{82}\\
&\ge (1 - \delta)^2 \ge 1 - 2\delta \tag{83}
\end{aligned}$$

So with probability at least $1 - \delta$,

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{B}}, \psi^*_{\mathcal{B}}(x), y)] \le{}& \kappa B \cdot d_{TV}(p_{\phi^*_{\mathcal{B}}}, p_{\phi^*_{\mathcal{A}}}) \tag{84}\\
&+ 3\kappa B \cdot \sqrt{\frac{1}{{n_{\mathcal{A}}}^2} \ln \frac{2N_{[]}(\mathcal{P}_{\mathcal{X}_{\mathcal{A}} \times \mathcal{S}}(\Phi_{\mathcal{A}}), \frac{1}{{n_{\mathcal{A}}}^2})}{\delta}} \tag{85}\\
&+ \kappa(\epsilon_{\mathcal{A}\mathcal{B}} + 2R_{m^2}(\mathcal{L}_{\text{CMD}} \circ \Phi_{\mathcal{B}}) + L\sqrt{\frac{2\ln(2/\delta)}{m^2}}) \tag{86}
\end{aligned}$$

By Chernoff bound, with probability at least $1 - \delta$, we have

$$\frac{1}{n_{\mathcal{B}}} \sum_{i=1}^{n_{\mathcal{B}}} \mathcal{L}(\hat{\phi}_{\mathcal{B}}, \psi^*_{\mathcal{B}}(x), y) - \mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{B}}, \psi^*_{\mathcal{B}}(x), y)] \le L\sqrt{\frac{2\ln(1/\delta)}{n_{\mathcal{B}}}} \tag{87}$$

Take $1 - \delta/2$ for equation 65 and equation 87, we get with probability $1 - \delta$,

$$\mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{B}}, \hat{\psi}_{\mathcal{B}}(x), y)] \le \epsilon_{\mathcal{B}} + \mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{B}}, \psi^*_{\mathcal{B}}(x), y)] + 2R_{n_{\mathcal{B}}}(\mathcal{L} \circ \Psi \circ \hat{\phi}_{\mathcal{B}}) + 2L\sqrt{\frac{2\ln(2/\delta)}{n_{\mathcal{B}}}} \tag{88}$$

Take $1 - \delta/2$ for equation 88 and equation 86, we get with probability $1 - \delta$,

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\hat{\psi}_{\mathcal{B}} \circ \hat{\phi}_{\mathcal{B}}(x), y)] \leq{} & \kappa B \cdot d_{TV}(\mathbb{P}_{\phi^*_{\mathcal{B}}}, \mathbb{P}_{\phi^*_{\mathcal{A}}}) + \kappa \epsilon_{\mathcal{A}\mathcal{B}} + \epsilon_{\mathcal{B}} && (89) \\
& + 2\kappa R_{m^2}(\mathcal{L}_{\text{CMD}} \circ \Phi_{\mathcal{B}}) + 2R_{n_{\mathcal{B}}}(\mathcal{L} \circ \Psi_{\mathcal{B}} \circ \hat{\phi}_{\mathcal{B}}) && (90) \\
& + 3\kappa B \cdot \sqrt{\frac{1}{{n_{\mathcal{A}}}^2} \ln \frac{4N_{[]}(\mathcal{P}_{\mathcal{X}_{\mathcal{A}} \times \mathcal{S}}(\Phi_{\mathcal{A}}), \frac{1}{{n_{\mathcal{A}}}^2})}{\delta}} + \kappa L \sqrt{\frac{2\ln(4/\delta)}{m^2}} + 2L\sqrt{\frac{2\ln(4/\delta)}{n_{\mathcal{B}}}} && (91)
\end{aligned}$$

□

B DISCUSSION ABOUT CMC LOSS

In order to introduce a similar bound for the CMC loss, we introduce a likelihood bound assumption,

Assumption B.1. *For any fixed ϕ, we have*

$$-\log \frac{p_{\phi_{\mathcal{A}},\phi_{\mathcal{B}}}(\boldsymbol{x})}{p_{\hat{\phi}_{\mathcal{A}},\phi_{\mathcal{B}}}(\boldsymbol{x})} \leq -(p_{\phi_{\mathcal{A}}}(\boldsymbol{x}) - p_{\hat{\phi}_{\mathcal{A}}}(\boldsymbol{x})) \log p_{\phi_{\mathcal{B}}}(\boldsymbol{x}) \tag{92}$$

where

$$p_{\phi_{\mathcal{A}},\phi_{\mathcal{B}}}(\boldsymbol{x}) = \frac{\exp(z_i^{\mathcal{A}} \cdot z_j^{\mathcal{B}}/\tau)}{\sum_t \exp(z_t^{\mathcal{A}} \cdot z_j^{\mathcal{B}}/\tau)}, \quad p_{\phi_{\mathcal{A}}}(\boldsymbol{x}) = \frac{\exp(z_i^{\mathcal{A}} \cdot z_j^{\mathcal{A}}/\tau)}{\sum_t \exp(z_t^{\mathcal{A}} \cdot z_j^{\mathcal{A}}/\tau)} \tag{93}$$

From the proofs above, we can find that changing the CMD loss to CMC loss does not affect the lemmas and theorems other than the final results Theorem 4.3. Thus, we just show that with the assumption B.1 we can get the same result as the Theorem 4.3 with CMC loss.

Proof. Noticing that the main difference for CMD and CMC losses are between equation 94 and equation 105. We only discuss the bounded process here and other derivations should be the same.

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\hat{\phi}_{\mathcal{B}}, \psi^*_{\mathcal{B}}(x), y)] &\leq \kappa \mathbb{E}[\mathbb{E}_{x'}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{B}}, (x, x'), s)]] && (94) \\
&= \kappa \mathbb{E}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{B}}, (x, x'), s)] \quad (\text{denote } \boldsymbol{x} = (x, x')) && (95) \\
&= \kappa(\mathbb{E}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{B}}, \boldsymbol{x}, s)] - \mathbb{E}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{A}}, \boldsymbol{x}, s)] && (96) \\
&\quad + \mathbb{E}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \phi^*_{\mathcal{A}}, \boldsymbol{x}, s)] - \mathbb{E}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)] && (97) \\
&\quad + \mathbb{E}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)]) && (98) \\
&= \kappa(\mathbb{E}[-\log \frac{p_{\phi^*_{\mathcal{B}},\hat{\phi}_{\mathcal{B}}}}{p_{\phi^*_{\mathcal{A}},\hat{\phi}_{\mathcal{B}}}}(\boldsymbol{x}, s)] + \mathbb{E}[-\log \frac{p_{\phi^*_{\mathcal{A}},\hat{\phi}_{\mathcal{B}}}}{p_{\hat{\phi}_{\mathcal{A}},\hat{\phi}_{\mathcal{B}}}}(\boldsymbol{x}, s)] && (99) \\
&\quad + \mathbb{E}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)]) && (100) \\
&\leq \kappa(\mathbb{E}[-p_{\phi^*_{\mathcal{B}}}(\boldsymbol{x}, s) \log p_{\hat{\phi}_{\mathcal{B}}}(\boldsymbol{x}, s) + p_{\phi^*_{\mathcal{A}}}(\boldsymbol{x}, s) \log p_{\hat{\phi}_{\mathcal{B}}}(\boldsymbol{x}, s)] && (101) \\
&\quad + \mathbb{E}[-p_{\phi^*_{\mathcal{A}}}(\boldsymbol{x}, s) \log p_{\hat{\phi}_{\mathcal{B}}}(\boldsymbol{x}, s) + p_{\hat{\phi}_{\mathcal{A}}}(\boldsymbol{x}, s) \log p_{\hat{\phi}_{\mathcal{B}}}(\boldsymbol{x}, s)] && (102) \\
&\quad + \mathbb{E}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)]) && (103) \\
&\leq \kappa B \cdot \Big(d_{TV}(\mathbb{P}_{\phi^*_{\mathcal{B}}}, \mathbb{P}_{\phi^*_{\mathcal{A}}}) + d_{TV}(\mathbb{P}_{\hat{\phi}_{\mathcal{A}}}, \mathbb{P}_{\phi^*_{\mathcal{A}}})\Big) && (104) \\
&\quad + \kappa \mathbb{E}[\mathcal{L}_{\text{CMC}}(\hat{\phi}_{\mathcal{B}}, \hat{\phi}_{\mathcal{A}}, \boldsymbol{x}, s)] && (105) \\
& && (106)
\end{aligned}$$

Then we get the same convergence bound of CMC loss as the CMD loss as shown in Theorem 4.3

□

Further improvement. The assumption B.1 in this paper is not trivial or prior to the analysis, further work to this research can focus on a better proof and result with CMC loss.

C DETAILED SETTINGS OF EXPERIMENTS

In this section, we give the detailed settings of datasets and training. In our experiments, all cross-modality distillation using a self-supervised learned ResNet on ImageNet. As mentioned in the paper, we only used the well-trained model provided by the official SimCLR with different structures of ResNet50, ResNet50(2x), and ResNet50(4x) but not using the ImageNet data in the distillation. To clarify the cross-modality distillation process, we give the dataset used for transferring and the detailed setting of the downstream task of each pair of modalities.

Training Dataset	Sketchy	TUBerlin	Sketchy-Eval
Train/Test Split	48,290	15,000/5,000	60,335/15,146
Optimizer	Adam	Adam	Adam
Optimizer Hyper-parameter	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)
Learning Rate Schedule	None	Multi-Step(60,70,80)	Multi-Step(60,70,80)
Learning Rate	1e-3	1e-3	1e-3
Epoch	100	100	100
Batch Size	64	64	64

Table 4: Details of image-sketch Distillation.

Since there are multiple sketches corresponding to one image in the Sketchy dataset, we consider all these pairs as positive pairs resulting in a total of 48290 training data. The train/split for TUBerlin and Sketchy-Eval just follows the typical setting used in Yu et al. (2017); Lin et al. (2020). Sketches in Sketchy-eval may have been trained without labels in distillation.

Training Dataset	NYU-Depth V2	NYU-Depth V2-Eval(Disjoint)
Train/Test Split	795	795/654
Optimizer	Adam	Adam
Optimizer Hyper-parameter	(0.9,0.999)	(0.9,0.999)
Learning Rate Schedule	None	Multi-Step(60,70,80)
Learning Rate	1e-3	1e-2
Epoch	100	100
Batch Size	16	16

Table 5: Details of image-depth map Distillation.

For the image-depth map task, we only use the training data in NYU-Depth V2 and also use the labeled version in downstream segmentation.

Training Dataset	VGGSound	VGGSound-Eval (Disjoint)
Train/Test Split	4,625	10,000/10,000
Optimizer	Adam	Adam
Optimizer Hyper-parameter	(0.9,0.999)	(0.9,0.999)
Learning Rate Schedule	None	Multi-Step(60,70,80)
Learning Rate	1e-3	1e-2
Epoch	100	100
Batch Size	16	16

Table 6: Details of video-audio Distillation.

In this case, we sample 4625 pairs of video and audio, translating the video into 12 frames and audio into spectrograms. A disjoint 10000 audio-only dataset is sampled to fine-tune downstream event classification where another 10000 are used for testing.