A. Missing Proofs in Section 3

In this section we provide all the missing proofs in Section 3. In what follows we will first restate the corresponding theorems for the ease of reading and then provide the detailed proofs.

Lemma 3.1. Let $\Sigma := \bigcup_{L\in\mathcal{L}} \Sigma_L$ and $\mathcal{D}_\Sigma$ be a language model over Σ^*. For any two string-to-string maps $f, f' : \Sigma^* \to \Sigma^*$, let $f_\sharp\mathcal{D}_\Sigma$ and $f'_\sharp\mathcal{D}_\Sigma$ be the corresponding pushforward distributions. Then $d_{\mathrm{TV}}(f_\sharp\mathcal{D}_\Sigma, f'_\sharp\mathcal{D}_\Sigma) \le \Pr_{\mathcal{D}_\Sigma}(f(X) \neq f'(X))$ where $X \sim \mathcal{D}_\Sigma$.

Proof. Note that the sample space Σ^* is countable. For any two distributions $\mathcal{P}$ and $\mathcal{Q}$ over Σ^*, it is a well-known fact that $d_{\mathrm{TV}}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2}\sum_{y\in\Sigma^*} |\mathcal{P}(y) - \mathcal{Q}(y)|$. Using this fact, we have:

$$
\begin{aligned}
d_{\mathrm{TV}}(f_\sharp\mathcal{D}, f'_\sharp\mathcal{D}) &= \frac{1}{2}\sum_{y\in\Sigma^*} \left|f_\sharp\mathcal{D}(y) - f'_\sharp\mathcal{D}(y)\right| \\
&= \frac{1}{2}\sum_{y\in\Sigma^*} \left|\Pr_{\mathcal{D}}(f(X) = y) - \Pr_{\mathcal{D}}(f'(X) = y)\right| \\
&= \frac{1}{2}\sum_{y\in\Sigma^*} |\mathbb{E}_{\mathcal{D}}[\mathbb{1}(f(X) = y)] - \mathbb{E}_{\mathcal{D}}[\mathbb{1}(f'(X) = y)]| \\
&\le \frac{1}{2}\sum_{y\in\Sigma^*} \mathbb{E}_{\mathcal{D}}\left[|\mathbb{1}(f(X) = y) - \mathbb{1}(f'(X) = y)|\right] \\
&= \frac{1}{2}\sum_{y\in\Sigma^*} \mathbb{E}_{\mathcal{D}}[\mathbb{1}(f(X) = y, f'(X) \neq y) + \mathbb{1}(f(X) \neq y, f'(X) = y)] \\
&= \frac{1}{2}\sum_{y\in\Sigma^*} \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}(f(X) = y, f'(X) \neq f(X))\right] + \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}(f'(X) = y, f'(X) \neq f(X))\right] \\
&= \sum_{y\in\Sigma^*} \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}(f(X) = y, f'(X) \neq f(X))\right] \\
&= \sum_{y\in\Sigma^*} \Pr_{\mathcal{D}}\left(f(X) = y, f'(X) \neq f(X)\right) \\
&= \Pr_{\mathcal{D}}(f(X) \neq f'(X)).
\end{aligned}
$$

The second equality holds by the definition of the pushforward distribution. The inequality on the fourth line holds due to the triangule inequality and the equality on the seventh line is due to the symmetry between $f(X)$ and $f'(X)$. The last equality holds by the total law of probability. ■

Theorem 3.1. (Lower bound, Two-to-One) Consider a setting of universal machine translation task with two source languages where $\Sigma^* = \Sigma^*_{L_0} \bigcup \Sigma^*_{L_1}$ and the target language is L. Let $g : \Sigma^* \to \mathcal{Z}$ be an ϵ-universal language mapping, then for any decoder $h : \mathcal{Z} \to \Sigma^*_L$, we have

$$
\begin{aligned}
&\mathrm{Err}^{L_0\to L}_{\mathcal{D}_0}(h\circ g) + \mathrm{Err}^{L_1\to L}_{\mathcal{D}_1}(h\circ g) \\
&\qquad\qquad \ge d_{\mathrm{TV}}(\mathcal{D}_{L_0,L}(L), \mathcal{D}_{L_1,L}(L)) - \epsilon.
\end{aligned} \tag{2}
$$

Proof of Theorem 3.1. First, realize that $d_{\mathrm{TV}}(\cdot,\cdot)$ is a distance metric, the following chain of triangle inequalities hold:

$$
\begin{aligned}
d_{\mathrm{TV}}(\mathcal{D}_{L_0,L}(L), \mathcal{D}_{L_1,L}(L)) \le\ & d_{\mathrm{TV}}(\mathcal{D}_{L_0,L}(L), (h\circ g)_\sharp\mathcal{D}_0) \\
&+ d_{\mathrm{TV}}((h\circ g)_\sharp\mathcal{D}_1, \mathcal{D}_{L_1,L}(L)) \\
&+ d_{\mathrm{TV}}((h\circ g)_\sharp\mathcal{D}_0, (h\circ g)_\sharp\mathcal{D}_1).
\end{aligned}
$$

Now by the assumption that g is an ϵ-universal language mapping and Corollary 3.1, the third term on the RHS of the above inequality, $d_{\mathrm{TV}}((h\circ g)_\sharp\mathcal{D}_0, (h\circ g)_\sharp\mathcal{D}_1)$, is upper bounded by ϵ. Furthermore, note that since the following equality holds:

$$
\mathcal{D}_{L_i,L}(L) = f^*_{L_i\to L\sharp}\mathcal{D}_i, \quad \forall i \in \{0,1\},
$$

we can further simplify the above inequality as

$$d_{\mathrm{TV}}(\mathcal{D}_{L_0,L}(L), \mathcal{D}_{L_1,L}(L)) \leq d_{\mathrm{TV}}(f^*_{L_0\to L}{}_\sharp \mathcal{D}_0, (h\circ g)_\sharp \mathcal{D}_0) + d_{\mathrm{TV}}((h\circ g)_\sharp \mathcal{D}_1, f^*_{L_1\to L}{}_\sharp \mathcal{D}_1) + \epsilon.$$

Now invoke Lemma 3.1 for $i \in \{0,1\}$ to upper bound the first two terms on the RHS, yielding:

$$d_{\mathrm{TV}}(f^*_{L_i\to L}{}_\sharp \mathcal{D}_i, (h\circ g)_\sharp \mathcal{D}_i) \leq \Pr_{\mathcal{D}_i}\big((h\circ g)(X) \neq f^*_{L_i\to L}(X)\big) = \mathrm{Err}^{L_i\to L}_{\mathcal{D}_i}(h\circ g).$$

A simple rearranging then completes the proof. ■

We now provide the proof of Theorem 3.2.

Theorem 3.2. (Lower bound, Many-to-Many) Consider a universal machine translation task where $\Sigma^* = \bigcup_{i\in[K]} \Sigma^*_{L_i}$. Let $\mathcal{D}_{L_i,L_k}$, $i,k \in [K]$ be the joint distribution of sentences (parallel corpus) in translating from L_i to L_k. If $g : \Sigma^* \to \mathcal{Z}$ be an ϵ-universal language mapping, then for any decoder $h : \mathcal{Z} \to \Sigma^*$, we have

$$\begin{aligned}
&\max_{i,k\in[K]} \mathrm{Err}^{L_i\to L_k}_{\mathcal{D}_{L_i,L_k}}(h\circ g) \geq \\
&\frac{1}{2}\max_{k\in[K]} \max_{i\neq j} d_{\mathrm{TV}}(\mathcal{D}_{L_i,L_k}(L_k), \mathcal{D}_{L_j,L_k}(L_k)) - \frac{\epsilon}{2}, \\
&\frac{1}{K^2}\sum_{i,k\in[K]} \mathrm{Err}^{L_i\to L_k}_{\mathcal{D}_{L_i,L_k}}(h\circ g) \geq \\
&\frac{1}{K^2(K-1)}\sum_{k\in[K]}\sum_{i<j} d_{\mathrm{TV}}(\mathcal{D}_{L_i,L_k}(L_k), \mathcal{D}_{L_j,L_k}(L_k)) - \frac{\epsilon}{2}.
\end{aligned}$$

Proof of Theorem 3.2. First let us fix a target language L_k. For each pair of source languages $L_i, L_j,\ i \neq j$ translating to L_k, applying Theorem 3.1 gives us:

$$\mathrm{Err}^{L_i\to L_k}_{\mathcal{D}_{L_i,L_k}}(h\circ g) + \mathrm{Err}^{L_j\to L_k}_{\mathcal{D}_{L_j,L_k}}(h\circ g) \geq d_{\mathrm{TV}}(\mathcal{D}_{L_i,L_k}(L_k), \mathcal{D}_{L_j,L_k}(L_k)) - \epsilon. \tag{10}$$

Now consider the pair of source languages (L_{i^*}, L_{j^*}) with the maximum $d_{\mathrm{TV}}(\mathcal{D}_{L_i,L_k}(L_k), \mathcal{D}_{L_j,L_k}(L_k))$:

$$\begin{aligned}
2\max_{i\in[K]} \mathrm{Err}^{L_i\to L_k}_{\mathcal{D}_{L_i,L_k}}(h\circ g) &\geq \mathrm{Err}^{L_{i^*}\to L_k}_{\mathcal{D}_{L_{i^*},L_k}}(h\circ g) + \mathrm{Err}^{L_{j^*}\to L_k}_{\mathcal{D}_{L_{j^*},L_k}}(h\circ g) \\
&\geq \max_{i\neq j}\ d_{\mathrm{TV}}(\mathcal{D}_{L_i,L_k}(L_k), \mathcal{D}_{L_j,L_k}(L_k)) - \epsilon.
\end{aligned} \tag{11}$$

Since the above lower bound (11) holds for any target language L_k, taking a maximum over the target languages yields:

$$2\max_{i,k\in[K]} \mathrm{Err}^{L_i\to L_k}_{\mathcal{D}_{L_i,L_k}}(h\circ g) \geq \max_{k\in[K]}\max_{i\neq j}\ d_{\mathrm{TV}}(\mathcal{D}_{L_i,L_k}(L_k), \mathcal{D}_{L_j,L_k}(L_k)) - \epsilon,$$

which completes the first part of the proof. For the second part, again, for a fixed target language L_k, to lower bound the average error, we apply the triangle inequality in (10) iteratively for all pairs $i < j$, yielding:

$$(K-1)\sum_{i\in[K]} \mathrm{Err}^{L_i\to L_k}_{\mathcal{D}_{L_i,L_k}}(h\circ g) \geq \sum_{i<j} d_{\mathrm{TV}}(\mathcal{D}_{L_i,L_k}(L_k), \mathcal{D}_{L_j,L_k}(L_k)) - \frac{K(K-1)}{2}\epsilon.$$

Dividing both sides by $K(K-1)$ gives the average translation error to L_k. Now summing over all the possible target language L_k yields:

$$\frac{1}{K^2}\sum_{i,k\in[K]} \mathrm{Err}^{L_i\to L_k}_{\mathcal{D}_{L_i,L_k}}(h\circ g) \geq \frac{1}{K^2(K-1)}\sum_{k\in[K]}\sum_{i<j} d_{\mathrm{TV}}(\mathcal{D}_{L_i,L_k}(L_k), \mathcal{D}_{L_j,L_k}(L_k)) - \frac{\epsilon}{2}.$$ ■

B. Missing Proofs in Section 4

In this section we provide all the missing proofs in Section 4. Again, in what follows we will first restate the corresponding theorems for the ease of reading and then provide the detailed proofs.

Lemma 4.1. If $S = \{(x_i, x_i')\}_{i=1}^n$ is sampled i.i.d. according to the encoder-decoder generative process, the following bound holds:

$$\Pr_{S\sim\mathcal{D}^n}\left(\sup_{f\in\mathcal{F}}|\varepsilon(f)-\widehat{\varepsilon}_S(f)|\geq\epsilon\right) \leq 2\mathcal{N}(\mathcal{F},\frac{\epsilon}{16M})\cdot\exp\left(\frac{-n\epsilon^2}{16M^4}\right).$$

Proof. For $f\in\mathcal{F}$, define $\ell_S(f) := \varepsilon(f)-\widehat{\varepsilon}_S(f)$ to be the generalization error of f on sample S. The first step is to prove the following inequality holds for $\forall f_1, f_2\in\mathcal{F}$ and any sample S:

$$|\ell_S(f_1)-\ell_S(f_2)|\leq 8M\cdot\|f_1-f_2\|_\infty.$$

In other words, $\ell_S(\cdot)$ is a Lipschitz function in $\mathcal{F}$ w.r.t. the ℓ_∞ norm. To see, by definition of the generalization error, we have

$$\begin{aligned}&|\ell_S(f_1)-\ell_S(f_2)|\\ &=|\varepsilon(f_1)-\widehat{\varepsilon}_S(f_1)-\varepsilon(f_2)+\widehat{\varepsilon}_S(f_2)|\\ &\leq|\varepsilon(f_1)-\varepsilon(f_2)|+|\widehat{\varepsilon}_S(f_1)-\widehat{\varepsilon}_S(f_2)|.\end{aligned}$$

To get the desired upper bound, it suffices for us to bound $|\varepsilon(f_1)-\varepsilon(f_2)|$ by $\|f_1-f_2\|_\infty$ and the same technique could be used to upper bound $|\widehat{\varepsilon}_S(f_1)-\widehat{\varepsilon}_S(f_2)|$ since the only difference lies in the measure where the expectation is taken over. We now proceed to upper bound $|\varepsilon(f_1)-\varepsilon(f_2)|$:

$$\begin{aligned}|\varepsilon(f_1)-\varepsilon(f_2)| &= \left|\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\|f_1(\mathbf{x})-\mathbf{x}'\|_2^2]-\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\|f_2(\mathbf{x})-\mathbf{x}'\|_2^2]\right|\\ &= \left|\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\|f_1(\mathbf{x})\|_2^2-\|f_2(\mathbf{x})\|_2^2-2\mathbf{x}'^T(f_1(\mathbf{x})-f_2(\mathbf{x}))]\right|\\ &\leq \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left|(f_1(\mathbf{x})-f_2(\mathbf{x}))^T(f_1(\mathbf{x})+f_2(\mathbf{x}))-2\mathbf{x}'^T(f_1(\mathbf{x})-f_2(\mathbf{x}))\right|\\ &\leq \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\left|(f_1(\mathbf{x})-f_2(\mathbf{x}))^T(f_1(\mathbf{x})+f_2(\mathbf{x}))\right|\right]+2\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\left|\mathbf{x}'^T(f_1(\mathbf{x})-f_2(\mathbf{x}))\right|\right]\\ &\leq \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\|f_1(\mathbf{x})-f_2(\mathbf{x})\|\cdot\|f_1(\mathbf{x})+f_2(\mathbf{x})\|\right]+2\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\|\mathbf{x}'\|\cdot\|f_1(\mathbf{x})-f_2(\mathbf{x})\|\right]\\ &\leq 2M\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\|f_1(\mathbf{x})-f_2(\mathbf{x})\|\right]+2M\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\|f_1(\mathbf{x})-f_2(\mathbf{x})\|\right]\\ &\leq 4M\|f_1-f_2\|_\infty.\end{aligned}$$

In the proof above, the first inequality holds due to the monotonicity property of integral. The second inequality holds by triangle inequality. The third one is due to Cauchy-Schwarz inequality. The fourth inequality holds by the assumption that $\forall f\in\mathcal{F}, \max_{x\in\mathcal{X}}\|f(\mathbf{x})\|\leq M$ and the identity mapping is in $\mathcal{F}$ so that $\|\mathbf{x}'\| = \|\mathrm{id}(\mathbf{x}')\|\leq\|\mathrm{id}(\cdot)\|_\infty\leq M$. The last one holds due to the monotonicity property of integral.

It is easy to see that the same argument could also be used to show that $|\widehat{\varepsilon}_S(f_1)-\widehat{\varepsilon}_S(f_2)|\leq 4M\|f_1-f_2\|_\infty$. Combine these two inequalities, we have

$$\begin{aligned}|\ell_S(f_1)-\ell_S(f_2)| &\leq |\varepsilon(f_1)-\varepsilon(f_2)|+|\widehat{\varepsilon}_S(f_1)-\widehat{\varepsilon}_S(f_2)|\\ &\leq 8M\|f_1-f_2\|_\infty.\end{aligned}$$

In the next step, we show that suppose $\mathcal{F}$ could be covered by k subsets $\mathcal{C}_1,\ldots,\mathcal{C}_k$, i.e., $\mathcal{F}=\cup_{i\in[k]}\mathcal{C}_i$. Then for any $\epsilon>0$, the following upper bound holds:

$$\Pr_{S\sim\mathcal{D}^n}\left(\sup_{f\in\mathcal{F}}|\ell_S(f)|\geq\epsilon\right)\leq\sum_{i\in[k]}\Pr_{S\sim\mathcal{D}^n}\left(\sup_{f\in\mathcal{C}_i}|\ell_S(f)|\geq\epsilon\right).$$

This follows from the union bound:

$$\begin{aligned}\Pr_{S\sim\mathcal{D}^n}\big(\sup_{f\in\mathcal{F}}|\ell_S(f)|\geq\epsilon\big) &= \Pr_{S\sim\mathcal{D}^n}\big(\bigcup_{i\in[k]}\sup_{f\in\mathcal{C}_i}|\ell_S(f)|\geq\epsilon\big)\\ &\leq \sum_{i\in[k]}\Pr_{S\sim\mathcal{D}^n}\big(\sup_{f\in\mathcal{C}_i}|\ell_S(f)|\geq\epsilon\big).\end{aligned}$$

Next, within each L_∞ ball $\mathcal{C}_i$ centered at f_i with radius $\frac{\epsilon}{16M}$ such that $\mathcal{F}\subseteq\cup_{i\in[k]}\mathcal{C}_i$, we bound each term in the above union bound as:

$$\Pr_{S\sim\mathcal{D}^n}\big(\sup_{f\in\mathcal{C}_i}|\ell_S(f)|\geq\epsilon\big)\leq\Pr_{S\sim\mathcal{D}^n}\big(|\ell_S(f_i)|\geq\epsilon/2\big).$$

To see this, realize that $\forall f\in\mathcal{C}_i$, we have $\|f-f_i\|_\infty\leq\epsilon/16M$, which implies

$$|\ell_S(f)-\ell_S(f_i)|\leq 8M\|f-f_i\|_\infty\leq\frac{\epsilon}{2}.$$

Hence we must have $|\ell_S(f_i)|\geq\epsilon/2$, otherwise $\sup_{f\in\mathcal{C}_i}|\ell_S(f)|<\epsilon$. This argument means that

$$\Pr_{S\sim\mathcal{D}^n}\big(\sup_{f\in\mathcal{C}_i}|\ell_S(f)|\geq\epsilon\big)\leq\Pr_{S\sim\mathcal{D}^n}\big(|\ell_S(f_i)|\geq\epsilon/2\big).$$

To finish the proof, we use the standard Hoeffding inequality to upper bound $\Pr_{S\sim\mathcal{D}^n}\big(|\ell_S(f_i)|\geq\epsilon/2\big)$ as follows:

$$\begin{aligned}\Pr_{S\sim\mathcal{D}^n}\big(|\ell_S(f_i)|\geq\epsilon/2\big) &= \Pr_{S\sim\mathcal{D}^n}\big(|\varepsilon(f_i)-\widehat{\varepsilon}_S(f_i)|\geq\epsilon/2\big)\\ &\leq 2\exp\left(-\frac{2n^2(\epsilon/2)^2}{n((2M)^2-0)^2}\right)\\ &= 2\exp\left(-\frac{n\epsilon^2}{16M^4}\right).\end{aligned}$$

Now combine everything together, we obtain the desired upper bound as stated in the lemma.

$$\Pr_{S\sim\mathcal{D}^n}\left(\sup_{f\in\mathcal{F}}|\varepsilon(f)-\widehat{\varepsilon}_S(f)|\geq\epsilon\right)\leq 2\mathcal{N}(\mathcal{F},\frac{\epsilon}{16M})\cdot\exp\left(\frac{-n\epsilon^2}{16M^4}\right).$$ ■

We next prove the generalization bound for a single pair of translation task:

Theorem 4.2. (Generalization, single task) Let S be a sample of size n according to our generative process. Then for any $0<\delta<1$, for any $f\in\mathcal{F}$, w.p. at least $1-\delta$, the following bound holds:

$$\varepsilon(f)\leq\widehat{\varepsilon}_S(f)+O\left(\sqrt{\frac{\log\mathcal{N}(\mathcal{F},\frac{\epsilon}{16M})+\log(1/\delta)}{n}}\right). \tag{7}$$

Proof. This is a direct corollary of Lemma 4.1 by setting the upper bound in Lemma 4.1 to be δ and solve for ϵ. ■

We now provide the proof sketch of Theorem 4.3. The main proof idea is exactly the same as the one we have in the deterministic setting, except that we replace the original definitions of errors and Lipschitzness with the generalized definitions under the randomized setting.

Theorem 4.3. (Sample complexity under generative model, randomized setting) Suppose H is connected and the trained $\{E_L\}_{L\in\mathcal{L}}$ satisfy

$$\forall L,L'\in H:\widehat{\varepsilon}_S(E_L,D_{L'})\leq\epsilon_{L,L'},$$

for $\epsilon_{L,L'}>0$. Furthermore, for $0<\delta<1$ suppose the number of sentences for each aligned corpora for each training pair (L,L') is $\Omega\left(\frac{1}{\epsilon^2_{L,L'}}\cdot\big(\log\mathcal{N}(\mathcal{F},\frac{\epsilon_{L,L'}}{16M})+\log(K/\delta)\big)\right)$. Then, with probability $1-\delta$, for any pair of languages $(L,L')\in\mathcal{L}\times\mathcal{L}$ and $L=L_1,L_2,\ldots,L_m=L'$ a path between L and L' in H, we have $\varepsilon(E_L,D_{L'})\leq 2\rho^2\sum_{k=1}^{m-1}\epsilon_{L_k,L_{k+1}}$.

Proof Sketch. The first step is prove the corresponding error concentration lemma using covering numbers as the one in Lemma 4.1. Again, due to the assumption that $\mathcal{F}$ is closed under composition, we have $D_{L'} \circ E_L \in \mathcal{F}$, hence it suffices if we could prove a uniform convergence bound for an arbitrary function $f \in \mathcal{F}$. To this end, for $f \in \mathcal{F}$, define $\ell_S(f) := \varepsilon(f) - \widehat{\varepsilon}_S(f)$ to be the generalization error of f on sample S. The first step is to prove the following inequality holds for $\forall f_1, f_2 \in \mathcal{F}$ and any sample S:

$$|\ell_S(f_1) - \ell_S(f_2)| \leq 8M \cdot \|f_1 - f_2\|_\infty.$$

In other words, $\ell_S(\cdot)$ is a Lipschitz function in $\mathcal{F}$ w.r.t. the ℓ_∞ norm. To see this, by definition of the generalization error, we have

$$|\ell_S(f_1) - \ell_S(f_2)| = |\varepsilon(f_1) - \widehat{\varepsilon}_S(f_1) - \varepsilon(f_2) + \widehat{\varepsilon}_S(f_2)| \leq |\varepsilon(f_1) - \varepsilon(f_2)| + |\widehat{\varepsilon}_S(f_1) - \widehat{\varepsilon}_S(f_2)|.$$

To get the desired upper bound, it suffices for us to bound $|\varepsilon(f_1) - \varepsilon(f_2)|$ by $\|f_1 - f_2\|_\infty$ and the same technique could be used to upper bound $|\widehat{\varepsilon}_S(f_1) - \widehat{\varepsilon}_S(f_2)|$ since the only difference lies in the measure where the expectation is taken over.

Before we proceed, in order to make the notation uncluttered, we first simplify $\varepsilon(f)$:

$$\varepsilon(f) = \mathbb{E}_{r,r'}\left[\|f - \mathbf{D}_{L'} \circ \mathbf{E}_L\|^2_{\ell_2(\mathbf{D}_{L\sharp}(\mathcal{D}\times\mathcal{D}_r))}\right].$$

Define $\mathbf{z} \sim \mathcal{D}$ to mean the sampling process of $(x, r, r') \sim \mathbf{D}_{L\sharp}(\mathcal{D} \times \mathcal{D}_r) \times D_r \times D_{r'}$, $\mathbf{x} := (x, r, r')$ and $\mathbf{x}' := \mathbf{D}_{L'}(\mathbf{E}_L(x, r'), r)$. Then

$$\begin{aligned}\varepsilon(f) &= \mathbb{E}_{r,r'}\left[\|f - \mathbf{D}_{L'} \circ \mathbf{E}_L\|^2_{\ell_2(\mathbf{D}_{L\sharp}(\mathcal{D}\times\mathcal{D}_r))}\right] \\ &= \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\|f(\mathbf{x}) - \mathbf{x}'\|_2^2].\end{aligned}$$

With the simplified notation, it is now clear that we essentially reduce the problem in the randomized setting to the original one in the deterministic setting. Hence by using exactly the same proof as the one of Lemma 4.1, we can obtain the following high probability bound:

$$\Pr\left(\sup_{f\in\mathcal{F}} |\varepsilon(f) - \widehat{\varepsilon}_S(f)| \geq \epsilon\right) \leq 2\mathcal{N}(\mathcal{F}, \frac{\epsilon}{16M}) \cdot \exp\left(\frac{-n\epsilon^2}{16M^4}\right).$$

As a direct corollary, a similar generalization bound for a single pair of translation task like the one in Theorem 4.2 also holds. To finish the proof, by the linearity of the expectation $\mathbb{E}_{r,r'}$, it is clear that exactly the same chaining argument in the proof of Theorem 4.1 could be used as well as the only thing we need to do is to take an additional expectation $\mathbb{E}_{r,r'}$ at the most outside level. ■