A THEORY FOR CONDITIONAL GENERATIVE MODELING ON MULTIPLE DATA SOURCES

Rongzhen Wang[1,2], **Yan Zhang**[3], **Chenyu Zheng**[1,2], **Chongxuan Li**[1,2]*, **Guoqiang Wu**[3]*
[1]Gaoling School of AI, Renmin University of China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods
[3]School of Software, Shandong University
{wangrz,cyzheng,chongxuanli}@ruc.edu.cn
yannzhang9@gmail.com, guoqiangwu@sdu.edu.cn

ABSTRACT

The success of large generative models has driven a paradigm shift, leveraging massive multi-source data to enhance model capabilities. However, the interaction among these sources remains theoretically underexplored. This paper takes a first step toward a rigorous analysis of multi-source training in conditional generative modeling, where each condition represents a distinct data source. Specifically, we establish a general distribution estimation error bound in average total variation distance for conditional maximum likelihood estimation (MLE) based on the bracketing number. Our result shows that when source distributions share similarity and the model is sufficiently expressive, multi-source training guarantees a sharper bound than single-source training. We further instantiate the general theory on conditional Gaussian estimation as an illustrative example. The result highlights that the number of sources and similarity among source distributions improve the advantage of multi-source training. Simulations and real-world experiments validate our findings. We hope this work inspires further theoretical understandings of multi-source training in generative modeling. Code is available at: https://github.com/ML-GSAI/Multi-Source-GM.

1 BACKGROUND

Large generative models have achieved remarkable success in generating realistic and complex outputs across natural language (Brown et al., 2020; Touvron et al., 2023) and computer vision (Rombach et al., 2022). A key factor behind their strong performance is the diverse and rich training data. For instance, large language models are trained on *heterogeneous* datasets comprising web content, books, and code (Brown et al., 2020; Hu et al., 2024), while image generation models benefit from vast datasets spanning various categories and aesthetic qualities (Peebles & Xie, 2023; Chen et al., 2024). Empirical evidence suggests that, under certain conditions, training on *multiple data sources* can enhance performance across all sources (Pires et al., 2019; Allen-Zhu & Li, 2024). Consequently, data mixture strategies have become an essential topic (Nguyen et al., 2022; Hu et al., 2024).

However, the theoretical underpinnings of this multi-source training paradigm remain poorly understood. This raises a fundamental question: *is it more effective to train separate models on individual data sources, or to train a single model using data from multiple sources?* In this paper, we take the first step toward a rigorous analysis of multi-source training, focusing on its impact on conditional generative models, where each condition represents a distinct data source. Our theoretical findings are validated through simulations and real-world experiments.

2 PROBLEM FORMULATION

We begin by mathematically describe conditional generative modeling via MLE with basic notations defined in Appendix A. The introduced multi-source setting abstracts practical scenarios where

*Correspondence to Chongxuan Li and Guoqiang Wu.

sources share common data structures while retaining unique characteristics. The single-source setting serves as a controlled baseline to assess the benefits of incorporating data from other sources.

2.1 DISTRIBUTIONS FOR MULTIPLE SOURCES

Let X denote the random variable for data (e.g., a natural image) in a data space $\mathcal{X}$, and Y denote the random variable for the source label in a label space $\mathcal{Y}$. Suppose there are K data sources (e.g., K categories of images), each corresponding to an unknown conditional distribution $p^*_{X|k}$ for $k \in [K]$. We assume $p^*_{X|k}$ is parameterized by a source-specific feature ϕ^*_k in parameter space Φ and a shared feature ψ^* in parameter space Ψ as $p^*_{X|k}(x|k) = p_{\phi^*_k,\psi^*}(x|k)$. The conditional distribution of X given $Y = y$ is consequently expressed as

$$p^*_{X|Y}(\boldsymbol{x}|y) = \prod_{k=1}^{K} \Big(p_{\phi^*_k,\psi^*}(\boldsymbol{x}|k) \Big)^{\mathbb{I}(y=k)}. \tag{1}$$

This compact representation provides convenience for subsequent discussions. We further assume the distribution of Y is known since the proportion of data from different sources is often manually designed in practice (Deng et al., 2009; Krizhevsky et al., 2009; Brown et al., 2020; Chen et al., 2024). The joint distribution of X and Y is then given by $p^*_{X,Y}(\boldsymbol{x}, y) = p^*_{X|Y}(\boldsymbol{x}|y)p^*_Y(y)$.

2.2 CONDITIONAL GENERATIVE MODELING

Given a dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ consisting of n independent and identically distributed (i.i.d.) data-label pairs sampled from the joint distribution $p^*_{X,Y}$, considering that a conditional generative model estimates $p^*_{X|Y}$ by MLE on S, where the conditional likelihood is defined as

$$\mathcal{L}_S(p_{X|Y}) \coloneqq \prod_{i=1}^{n} p_{X|Y}(\boldsymbol{x}_i|y_i). \tag{2}$$

Under multi-source training, the conditional distribution space is given by

$$\mathcal{P}^{\text{multi}}_{X|Y} \coloneqq \Big\{ p^{\text{multi}}_{X|Y}(\boldsymbol{x}|y) = \textstyle\prod_{k=1}^{K} \big(p_{\phi_k,\psi}(\boldsymbol{x}|k) \big)^{\mathbb{I}(y=k)} : \phi_k \in \Phi, \psi \in \Psi \Big\},$$

and the corresponding estimator is

$$\hat{p}^{\text{multi}}_{X|Y} = \operatorname*{arg\,max}_{p^{\text{multi}}_{X|Y} \in \mathcal{P}^{\text{multi}}_{X|Y}} \mathcal{L}_S(p^{\text{multi}}_{X|Y}). \tag{3}$$

Here, we adopt the realizable assumption that true parameters satisfy $\phi^*_k \in \Phi$ and $\psi^* \in \Psi$, which allows the estimation error analysis to focus on the generalization property of the distribution space.

Under single-source training, we train K conditional generative models for each source using data exclusively from the corresponding source. For any particular source k, denoting $S_k \coloneqq \{(\boldsymbol{x}_i, y_i) \in S | y_i = k\} = \{\boldsymbol{x}^k_j, k\}_{j=1}^{n_k}$, the corresponding generative model estimate $p^*_{X|k}$ by maximizing the conditional likelihood on S_k as $\hat{p}^{\text{single}}_{X|k} = \arg\max_{p^{\text{single}}_{X|k} \in \mathcal{P}^{\text{single}}_{X|k}} \mathcal{L}_{S_k}(p^{\text{single}}_{X|k})$, where $\mathcal{L}_{S_k}(p_{X|k}) \coloneqq \prod_{j=1}^{n_k} p_{X|k}(\boldsymbol{x}^k_j|k)$ and $\mathcal{P}^{\text{single}}_{X|k} \coloneqq \big\{ p_{\phi_k,\psi_k}(\boldsymbol{x}|k) : \phi_k \in \Phi, \psi_k \in \Psi \big\}$. Separately maximizing these K objectives is equivalent to maximizing L_S in conditional distribution space $\mathcal{P}^{\text{single}}_{X|Y} \coloneqq \Big\{ p^{\text{single}}_{X|Y}(\boldsymbol{x}|y) = \prod_{k=1}^{K} \big(p_{\phi_k,\psi_k}(\boldsymbol{x}|k) \big)^{\mathbb{I}(y=k)} : \phi_k \in \Phi, \psi_k \in \Psi \Big\}$. Therefore, the estimator of $p^*_{X|Y}$ under single-source training is

$$\hat{p}^{\text{single}}_{X|Y} = \operatorname*{arg\,max}_{p^{\text{single}}_{X|Y} \in \mathcal{P}^{\text{single}}_{X|Y}} \mathcal{L}_S(p^{\text{single}}_{X|Y}). \tag{4}$$

We measure the accuracy of conditional distribution estimation by the average TV distance between the estimated and true conditional distributions, referred to as the *average TV error*, defined as:

$$\mathcal{R}_{\overline{\text{TV}}}(\hat{p}_{X|Y}) \coloneqq \mathbb{E}_Y \Big[\text{TV}(\hat{p}_{X|Y}, p^*_{X|Y}) \Big], \tag{5}$$

where the TV distance is given by $\text{TV}(\hat{p}_{X|y}, p^*_{X|y}) = \frac{1}{2} \int_{\mathcal{X}} |\hat{p}_{X|Y}(\boldsymbol{x}|y) - p^*_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x}$.

3 PROVABLE ADVANTAGE OF MULTI-SOURCE TRAINING

In this section, we first establish a general upper bound on the average TV error for conditional MLE and then prove a guaranteed advantage for multi-source training. This analysis extends classical MLE error bounds (Wong & Shen, 1995; Ge et al., 2024) to the conditional setting by introducing the *upper bracketing number* to quantify the complexity of conditional distribution space and modify the proofs to handle conditional MLE. Detailed discussions and definitions are deferred to Appendix B.

Theorem 3.1 (Average TV error bound for conditional MLE, proof in Appendix B.2.)**.** *Given a dataset S of size n that i.i.d. sampled from $p^*_{X,Y}$, let $\hat{p}_{X|Y}$ be the maximizer of $L_S(p_{X|Y})$ defined in Equation* (2) *in conditional distribution space $\mathcal{P}_{X|Y}$. Suppose the real conditional distribution $p^*_{X|Y}$ is contained in $\mathcal{P}_{X|Y}$. Then, for any $0 < \delta \leq 1/2$, it holds with probability at least $1-\delta$ that*

$$\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}_{X|Y}) \leq 3\sqrt{\tfrac{1}{n}\left(\log \mathcal{N}_{[]}\left(\tfrac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right) + \log \tfrac{1}{\delta}\right)}.$$

Here, $\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)$ denotes the $\frac{1}{n}$-upper bracketing number of $\mathcal{P}_{X|Y}$ w.r.t. $L^1(\mathcal{X})$ as defined in Definition B.1. Notably, as formulated in Section 2, Theorem 3.1 is applicable to both multi-source and single-source training. The following proposition further shows that multi-source training reduces the bracketing number of its distribution space through source similarity.

Proposition 3.2 (Multi-source training reducing complexity, proof in Appendix B.3.)**.** *Let $\mathcal{P}^{\mathrm{multi}}_{X|Y}$ and $\mathcal{P}^{\mathrm{single}}_{X|Y}$ be as defined in Section 2. Then, for any $\epsilon > 0$ and $1 \leq \mathsf{p} \leq \infty$, we have*

$$\mathcal{N}_{[]}\left(\epsilon; \mathcal{P}^{\mathrm{multi}}_{X|Y}, L^{\mathsf{p}}(\mathcal{X})\right) \leq \mathcal{N}_{[]}\left(\epsilon; \mathcal{P}^{\mathrm{single}}_{X|Y}, L^{\mathsf{p}}(\mathcal{X})\right).$$

Combining Theorem 3.1 and Proposition 3.2, we conclude that when source distributions exhibit parametric similarity and the realizable assumption is satisfied, multi-source training can enjoy a sharper estimation guarantee than single-source training. To clearly illustrate this advantage, the next section presents a concrete example by explicitly measuring the corresponding bracketing numbers.

4 INSTANTIATION ON CONDITIONAL GAUSSIAN ESTIMATION

Now, we instantiate the general theory using Gaussian models as employed in extensive work (Montanari & Saeed, 2022; Zheng et al., 2023; Dandi et al., 2024), which offer a simple yet insightful example and enable analytically tractable simulations under our theoretical assumptions.

Suppose each conditional distribution is a d-dimensional standard Gaussian distribution, i.e., $X|k \sim \mathcal{N}(\boldsymbol{\mu}^*_k, \boldsymbol{I}_d) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}^*_k\|_2^2}$ with a mean vector $\boldsymbol{\mu}^*_k$ and an identity covariance matrix $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ for all $k \in [K]$. We assume each $\boldsymbol{\mu}^*_k$ has two parts: the first d_1 entries $\boldsymbol{\mu}^*_k[1:d_1]$ represent the source-specific feature which is potentially different for each source, and the remaining entries $\boldsymbol{\mu}^*_k[d_1+1:d]$ represent the shared feature which is identical across all sources. Corresponding to the general formulation in Section 2, we denote $\phi_k \coloneqq \boldsymbol{\mu}^*_k[1:d_1], \psi \coloneqq \boldsymbol{\mu}^*_1[d_1+1:d] = \cdots = \boldsymbol{\mu}^*_K[d_1+1:d]$, and then the conditional distribution is parameterized as

$$p_{\phi_k,\psi}(\boldsymbol{x}|k) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|\boldsymbol{x}-(\phi_k,\psi)\|_2^2}. \tag{6}$$

Under this formulation, multi-source training leads to the following result.

Theorem 4.1 (Average TV error bound for conditional Gaussian parametric estimation under multi--source training, proof in Appendix C.2)**.** *Let $\hat{p}^{\mathrm{multi}}_{X|Y}$ be the likelihood maximizer defined in Equation* (3) *given $\mathcal{P}^{\mathrm{multi}}_{X|Y}$ with conditional distributions as in Equation* (6)*. Suppose $\Phi = [-B, B]^{d_1}$, $\Psi = [-B, B]^{d-d_1}$ with constant $B > 0$, and $\phi^*_k \in \Phi$, $\psi^* \in \Psi$. Then, for any $0 < \delta \leq 1/2$, it holds with probability at least $1-\delta$ that $\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{multi}}_{X|Y}) = \tilde{\mathcal{O}}\left(\sqrt{(K-1)d_1 + d/n}\right)$.*

In contrast, single-source training results in $\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{single}}_{X|Y}) = \tilde{\mathcal{O}}\left(\sqrt{Kd/n}\right)$ provided in Theorem C.2.

The advantage of multi-source learning can be quantified by the ratio of these error bounds:

$$\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{multi}}_{X|Y})/\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{single}}_{X|Y}) = \sqrt{\frac{(K-1)d_1+d}{Kd}} = \sqrt{1-\frac{K-1}{K}\beta_{\mathrm{sim}}},$$

where $\beta_{\mathrm{sim}} \coloneqq (d-d_1)/d$ measures the proportion of shared dimensions. As K increases from 1 to ∞, the ratio decreases from 1 to $1-\beta_{\mathrm{sim}}$, and as β_{sim} increases from 0 (completely dissimilar source distributions) to 1 (completely identical source distributions), it decreases from 1 to $\sqrt{1/K}$, reflecting a transition from no asymptotic gain to a constant improvement. This highlights that the number of sources and distribution similarity enhance the advantage of multi-source training.

5 Experiments

Due to the page limit, we only report the experimental results in Figure 1 and Figure 2. Detailed settings and additional interpretations are provided in Appendix D.

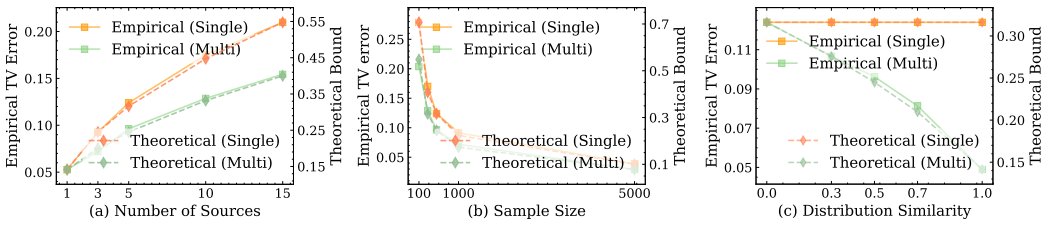

Figure 1: Simulation results for conditional Gaussian estimation. Empirical values (solid lines) correspond to the left vertical axis, while theoretical values (dashed lines) correspond to the right. Single-source results are shown in orange, and multi-source results in green.

Table 1: Average FID for single-source and multi-source training. Under different amounts of classes K, similarity level Sim, and per-class sample size N, multi-source training generally achieves lower average FID than that of single-source training.

N	Sim	K	Avg. FID ↓ (Single)	Avg. FID ↓ (Multi)
500	1	3	30.15	**29.82**
		10	30.16	**29.36**
	2	3	32.87	**31.16**
		10	29.96	**28.83**
1000	1	3	27.76	**26.86**
		10	27.46	**25.01**
	2	3	30.20	**28.31**
		10	29.60	**26.70**

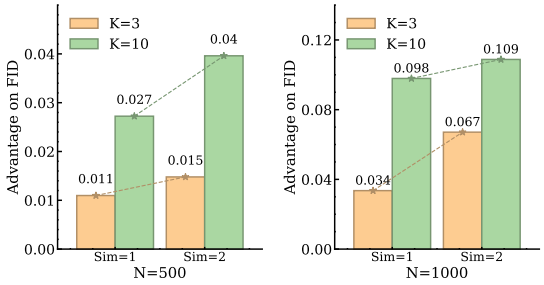

Figure 2: Relative advantage of multi-source training. For any fixed similarity level Sim and per-class sample size N, the relative advantage of multi-sources training with a larger K is larger than that with a smaller K. For any fixed K and N, the relative advantage of multi-source training with a larger distribution similarity is larger than that with a smaller distribution similarity (as shown through the dashed lines).

6 Conlusion

This paper provides the first attempt to rigorously analyze the conditional generative modeling on multiple data sources from a distribution estimation perspective. In particular, we establish a general estimation error bound in average TV distance under the realizable assumption based on the bracketing number of the conditional distribution space. When source distributions share parametric similarity, multi-source training has a provable advantage against single-source training by reducing the bracketing number. We further instantiate the general theory on conditional Gaussian estimation to obtain concrete error bounds. The result shows that the number of data sources and the similarity between source distributions enhance the advantage of multi-source training guarantee.

ACKNOWLEDGMENTS

This work was supported by NSF of China (62206159); Beijing Nova Program (20220484044); Beijing Natural Science Foundation (L247030); Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (22XNKJ13); the Natural Science Foundation of Shandong Province (ZR2022QF117), the Fundamental Research Funds of Shandong University. The work was partially done at the Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education. G. Wu was also sponsored by the TaiShan Scholars Program (NO.tsqn202306051).

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Mike Bostock. Imagenet hierarchy, 2018. URL `https://observablehq.com/@mbostock/imagenet-hierarchy`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Universality laws for gaussian mixtures in generalized linear models. *Advances in Neural Information Processing Systems*, 36, 2024.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet summary and statistics, 2010. URL `https://tex.stackexchange.com/questions/3587/how-can-i-use-bibtex-to-cite-a-web-page`.

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Sara A Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies. *CoRR*, abs/2404.06395, 2024.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pp. 4310–4312. PMLR, 2022.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of CLIP. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182. IEEE, 2023.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4996–5001. Association for Computational Linguistics, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

Jon A Wellner. Empirical processes in statistics: Methods, examples, further problems, 2002.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, pp. 339–362, 1995.

Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. Toward understanding generative data augmentation. *Advances in neural information processing systems*, 36:54046–54060, 2023.

A ELEMENTARY NOTATIONS

Scalars, vectors, and matrices are denoted by lowercase letters (e.g., a), lowercase boldface letters (e.g., $\boldsymbol{a}$), and uppercase boldface letters (e.g., $\boldsymbol{A}$). We use $\boldsymbol{a}[m]$ to denote the m-th entry of vector $\boldsymbol{a}$, and $\boldsymbol{A}[m,:]$, $\boldsymbol{A}[:,n]$, and $\boldsymbol{A}[m,n]$ to denote the m-th row, the n-th column, and the entry at the m-th row and the n-th column of $\boldsymbol{A}$. $(\boldsymbol{a},\boldsymbol{b})$ denotes the concatenation of $\boldsymbol{a}$ and $\boldsymbol{b}$. We denote $[n] \coloneqq \{1,\ldots,n\}$ for any $n \in \mathbb{N}$ and $a \vee b$ as $\max\{a,b\}$. For any measurable scalar function $f(\boldsymbol{x})$ on domain $\mathcal{X}$ and real number $1 \leq \mathsf{p} \leq \infty$, its $L^{\mathsf{p}}(\mathcal{X})$-norm is defined as $\|f(\boldsymbol{x})\|_{L^{\mathsf{p}}(\mathcal{X})} \coloneqq (\int_{\mathcal{X}} |f(\boldsymbol{x})|^{\mathsf{p}}\, d\boldsymbol{x})^{\frac{1}{\mathsf{p}}}$. When $\mathsf{p}=\infty$, $\|f(\boldsymbol{x})\|_{L^{\infty}(\mathcal{X})} = \sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x})|$. $\mathbb{I}(\cdot)$ denotes the indicator function. Notation $a_n = \tilde{\mathcal{O}}(b_n)$ indicates a_n is asymptotically bounded above by b_n up to logarithmic factors.

B PROOFS FOR SECTION 3

B.1 COMPLEXITY OF THE CONDITIONAL DISTRIBUTION SPACE

We begin by introducing an extended notion of the bracketing number as follows.

Definition B.1 (ϵ-upper bracketing number for conditional distribution space)**.** Let ϵ be a real number that $\epsilon > 0$ and p be an integer that $1 \leq \mathsf{p} \leq \infty$. An ϵ-upper bracket of a conditional distribution space $\mathcal{P}_{X|Y}$ with respect to $L^{\mathsf{p}}(\mathcal{X})$ is a finite function set $\mathcal{B}$ such that for any $p_{X|Y} \in \mathcal{P}_{X|Y}$, there exists some $p' \in \mathcal{B}$ such that given any $y \in \mathcal{Y}$, it holds

$$\forall \boldsymbol{x} \in \mathcal{X} : p'(\boldsymbol{x},y) \geq p_{X|Y}(\boldsymbol{x}|y), \text{ and } \|p'(\cdot,y) - p_{X|Y}(\cdot|y)\|_{L^{\mathsf{p}}(\mathcal{X})} \leq \epsilon.$$

The ϵ-upper bracketing number $\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^{\mathsf{p}}(\mathcal{X})\Big)$ is the cardinality of the smallest ϵ-upper bracket.

This notion quantifies the minimal set of functions needed to upper bound every conditional distribution within a small margin, reducing error analysis from an infinite to a finite function class. Unlike traditional bracketing numbers for unconditional distributions p_X using two-sided brackets (Wellner, 2002), this extension employs one-sided upper brackets (Ge et al., 2024) and requires uniform coverage across y for all conditional distributions tailored for our setting.

B.2 PROOF OF THEOREM 3.1

Proof of Theorem 3.1. Classical approaches investigate distribution estimation for MLE in Hellinger distance based on the bracketing number and the uniform law of large numbers from empirical process theory (Wong & Shen, 1995; Geer, 2000), which yields high-probability bounds of similar order as Theorem 3.1. Ge et al. (2024) extend the analysis to derive TV error bound under the realizable assumption.

We further adapt the techniques in Ge et al. (2024) to conditional generative modeling by introducing the upper bracketing number to quantify the complexity of conditional distribution space in Definition B.1 and modify the proofs to handle conditional MLE. The formal proof is presented below. Notably, the theorem applies to both discrete and continuous random variables, while we use integration notation in the proof for generality.

In the following, we first present an elementary inequality (in Equation (9)) which serves as a toolkit for the subsequent derivations. Then we decompose the TV distance and derive its complexity-based upper bound (in Equation (11)) using the former inequality. Finally, after specifying certain constants in this upper bound, a clearer order w.r.t. n is revealed (in Equation (12)).

Intermediate result induced by union bound. Let ϵ be a real number that $\epsilon > 0$ and p be an integer that $1 \leq \mathsf{p} \leq \infty$. Let $\mathcal{B}$ be an ϵ-upper bracket of $\mathcal{P}_{X|Y}$ w.r.t. $L^1(\mathcal{X})$ such that $|\mathcal{B}| = \mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)$.

According to the minimum cardinality requirement, we obtain a proposition of $\mathcal{B}$ that: for any $p' \in \mathcal{B}$, $p'(\boldsymbol{x},y) \geq 0$ on $\mathcal{X}\times\mathcal{Y}$. Let's first consider $\prod_{i=1}^{n} \sqrt{\frac{p'(\boldsymbol{x}_i,y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}}$ as a random variable on S, where

we suppose $p^*_{X,Y}(\boldsymbol{x}_i, y_i) > 0$ since $(\boldsymbol{x}_i, y_i)$ are sampled from $p^*_{X,Y}$ and thus $p^*_{X|Y}(\boldsymbol{x}_i|y_i) \neq 0$. By applying the *Markov inequality*, we have: given any $0 < \delta' < 1$,

$$\Pr_S\left(\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}} \geq \frac{1}{\delta'}\mathbb{E}_S\left[\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}}\right]\right) \leq \delta'. \tag{7}$$

Applying the *union bound* on all $p' \in \mathcal{B}$, we further have:

$$\begin{aligned}
&\Pr_S\left(\forall p' \in \mathcal{B}, \prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}} < \frac{1}{\delta'}\mathbb{E}_S\left[\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}}\right]\right) \\
=&1 - \Pr_S\left(\exists p' \in \mathcal{B}, \prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}} \geq \frac{1}{\delta'}\mathbb{E}_S\left[\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}}\right]\right) \\
=&1 - \Pr_S\left(\bigcup_{p' \in \mathcal{B}}\left\{\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}} \geq \frac{1}{\delta'}\mathbb{E}_S\left[\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}}\right]\right\}\right) \\
\geq&1 - \sum_{p' \in \mathcal{B}} \Pr_S\left(\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}} \geq \frac{1}{\delta'}\mathbb{E}_S\left[\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}}\right]\right) && \text{(by \textit{union bound})} \\
\geq&1 - \mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)\delta'. && \text{(by Equation (7))}
\end{aligned}$$

By denoting that $\delta := \mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)\delta'$, we have: it holds with probability at least $1 - \delta$ that for all $p' \in \mathcal{B}$,

$$\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}} < \frac{\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)}{\delta}\mathbb{E}_S\left[\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}}\right].$$

Taking logarithms at both sides, we have

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^n \log\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)} &\leq \log\mathbb{E}_S\left[\prod_{i=1}^n \sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}}\right] + \log\frac{\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)}{\delta} \\
&= \log\prod_{i=1}^n \mathbb{E}_{(\boldsymbol{x}_i, y_i)\sim p^*_{X,Y}}\left[\sqrt{\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)}}\right] + \log\frac{\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)}{\delta} \\
&\qquad (\{x_i\}_{i=1}^n \text{ are i.i.d. sampled from } p^*_X) \\
&= n\log\mathbb{E}_{X,Y}\left[\sqrt{\frac{p'(\boldsymbol{x}, y)}{p^*_{X|Y}(\boldsymbol{x}|y)}}\right] + \log\frac{\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)}{\delta} \\
&= n\log\mathbb{E}_Y\left[\mathbb{E}_{X|Y}\left[\sqrt{\frac{p'(\boldsymbol{x}, y)}{p^*_{X|Y}(\boldsymbol{x}|y)}}\right]\right] + \log\frac{\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)}{\delta} \\
&= n\log\mathbb{E}_Y\left[\int_{\mathcal{X}} p^*_{X|Y}(\boldsymbol{x}|y)\sqrt{\frac{p'(\boldsymbol{x}, y)}{p^*_{X|Y}(\boldsymbol{x}|y)}}d\boldsymbol{x}\right] + \log\frac{\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)}{\delta} \\
&= n\log\mathbb{E}_Y\left[\int_{\mathcal{X}} \sqrt{p'(\boldsymbol{x}, y)p^*_{X|Y}(\boldsymbol{x}|y)}d\boldsymbol{x}\right] + \log\frac{\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\Big)}{\delta}.
\end{aligned}$$

As $\log x \leq x - 1$ for all $x > 0$, the inequality can be further transformed into

$$\frac{1}{2}\sum_{i=1}^{n}\log\frac{p'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)} \leq n\left(\mathbb{E}_Y\left[\int_{\mathcal{X}}\sqrt{p'(\boldsymbol{x}, y)p^*_{X|Y}(\boldsymbol{x}|y)}d\boldsymbol{x}\right] - 1\right) + \log\frac{\mathcal{N}_{[]}\left(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta}. \tag{8}$$

Elementary inequality for MLE estimators. Since the real conditional distribution $p^*_{X|Y}$ is in $\mathcal{P}_{X|Y}$, for the likelihood maximizers $\hat{p}_{X|Y} \in \mathcal{P}_{X|Y}$, we have $L_S(\hat{p}_{X|Y}) = \prod_{i=1}^n \hat{p}_{X|Y}(\boldsymbol{x}_i|y_i) \geq L_S(p^*_{X|Y}) = \prod_{i=1}^n p^*_{X|Y}(\boldsymbol{x}_i|y_i)$, and thus $\frac{1}{2}\sum_{i=1}^n \log\frac{\hat{p}_{X|Y}(\boldsymbol{x}_i|y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)} = \frac{1}{2}\log\frac{\prod_{i=1}^n \hat{p}_{X|Y}(\boldsymbol{x}_i|y_i)}{\prod_{i=1}^n p^*_{X|Y}(\boldsymbol{x}_i|y_i)} \geq \frac{1}{2}\log 1 = 0$. According to the definition of upper bracketing number, there exists some $\hat{p}' \in \mathcal{B}$ such that given any $y \in \mathcal{Y}$, it holds that: (i) $\forall x \in \mathcal{X}, \hat{p}'(\boldsymbol{x}, y) \geq \hat{p}_{X|Y}(\boldsymbol{x}|y)$, and (ii) $\|\hat{p}'(\cdot, y) - \hat{p}_{X|Y}(\cdot|y)\|_{L^1(\mathcal{X})} = \int_{\mathcal{X}}|\hat{p}'(\boldsymbol{x}, y) - \hat{p}_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x} \leq \epsilon$. Applying (i), we have:

$$\frac{1}{2}\sum_{i=1}^{n}\log\frac{\hat{p}'(\boldsymbol{x}_i, y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)} \geq \frac{1}{2}\sum_{i=1}^{n}\log\frac{\hat{p}_{X|Y}(\boldsymbol{x}_i|y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)} \geq 0.$$

Combining this with Equation (8) and rearranging the terms, we have: it holds with at least probability $1 - \delta$ that

$$1 - \mathbb{E}_Y\left[\int_{\mathcal{X}}\sqrt{p'(\boldsymbol{x}, y)p^*_{X|Y}(\boldsymbol{x}|y)}d\boldsymbol{x}\right] \leq \frac{1}{n}\log\frac{\mathcal{N}_{[]}\left(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta}. \tag{9}$$

This serves as an elementary toolkit for deriving the subsequent upper bounds.

Decomposing the square of the TV distance. Recalling that $\mathrm{TV}(\hat{p}_{X|Y}, p^*_{X|Y}) = \frac{1}{2}\int_{\mathcal{X}}|\hat{p}_{X|Y}(\boldsymbol{x}|y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}$, we will decompose its square and then bound each term sequentially. First, we use the above $\hat{p}'(\boldsymbol{x}, y)$ as an intermediate term to decompose the square of $2\mathrm{TV}(\hat{p}_{X|Y}, p^*_{X|Y})$ into parts that can be effectively upper bounded:

$$\begin{aligned}
&\left(2\mathrm{TV}(\hat{p}_{X|Y}, p^*_{X|Y})\right)^2 = \left(\int_{\mathcal{X}}|\hat{p}_{X|Y}(\boldsymbol{x}|y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}\right)^2 \\
=&\underbrace{\left(\int_{\mathcal{X}}|\hat{p}_{X|Y}(\boldsymbol{x}|y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}\right)^2 - \left(\int_{\mathcal{X}}|\hat{p}'(\boldsymbol{x}, y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}\right)^2}_{\text{(I)}} + \\
&\underbrace{\left(\int_{\mathcal{X}}|\hat{p}'(\boldsymbol{x}, y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}\right)^2}_{\text{(II)}}.
\end{aligned}$$

For (I), we have

$$\begin{aligned}
&\left(\int_{\mathcal{X}}|\hat{p}_{X|Y}(\boldsymbol{x}|y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}\right)^2 - \left(\int_{\mathcal{X}}|\hat{p}'(\boldsymbol{x}, y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}\right)^2 \\
=&\left(\int_{\mathcal{X}}|\hat{p}_{X|Y}(\boldsymbol{x}|y) - p^*_{X|Y}(\boldsymbol{x}|y)| + |\hat{p}'(\boldsymbol{x}, y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}\right) \\
&\left(\int_{\mathcal{X}}|\hat{p}_{X|Y}(\boldsymbol{x}|y) - p^*_{X|Y}(\boldsymbol{x}|y)| - |\hat{p}'(\boldsymbol{x}, y) - p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}\right) \\
\leq&\left(\int_{\mathcal{X}}|\hat{p}_{X|Y}(\boldsymbol{x}|y)| + |p^*_{X|Y}(\boldsymbol{x}|y)| + |\hat{p}'(\boldsymbol{x}, y) - \hat{p}_{X|Y}(\boldsymbol{x}|y)| + |\hat{p}_{X|Y}(\boldsymbol{x}|y)| + |p^*_{X|Y}(\boldsymbol{x}|y)|d\boldsymbol{x}\right) \\
&\left(\int_{\mathcal{X}}|\hat{p}_{X|Y}(\boldsymbol{x}|y) - \hat{p}'(\boldsymbol{x}, y)|d\boldsymbol{x}\right) \\
\leq&(\epsilon + 4)\epsilon.
\end{aligned}$$

The first inequality holds for the *triangle inequality* $|a+b| \leq |a| + |b|$ and the *reverse triangle inequality* $||a| - |b|| \leq |a-b|$. The second inequality holds for the normalization property of conditional distributions ($\int_{\mathcal{X}} |\hat{p}_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x}$ and $\int_{\mathcal{X}} |p^*_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x}$ equal 1) and the property of the ϵ-upper bracket ($\int_{\mathcal{X}} |\hat{p}'(\boldsymbol{x}, y) - \hat{p}_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x} \leq \epsilon$).

For (II), we have

$$
\begin{aligned}
&\left(\int_{\mathcal{X}} |\hat{p}'(\boldsymbol{x}, y) - p^*_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x}\right)^2 \\
\leq &\left(\int_{\mathcal{X}} \left(\sqrt{\hat{p}'(\boldsymbol{x}, y)} + \sqrt{p^*_{X|Y}(\boldsymbol{x}|y)}\right)^2 dx\right) \left(\int_{\mathcal{X}} \left(\sqrt{\hat{p}'(\boldsymbol{x}, y)} - \sqrt{p^*_{X|Y}(\boldsymbol{x}|y)}\right)^2 dx\right) \\
&\qquad \text{(by } \textit{Cauchy–Schwarz inequality}\text{)} \\
\leq &\left(\int_{\mathcal{X}} 2\Big(\hat{p}'(\boldsymbol{x}, y) + p^*_{X|Y}(\boldsymbol{x}|y)\Big) d\boldsymbol{x}\right) \left(\int_{\mathcal{X}} \hat{p}'(\boldsymbol{x}, y) + p^*_{X|Y}(\boldsymbol{x}|y) - 2\sqrt{\hat{p}'(\boldsymbol{x}, y) p^*_{X|Y}(\boldsymbol{x}|y)} d\boldsymbol{x}\right) \\
&\qquad \text{(by } (a+b)^2 \leq 2(a^2+b^2)\text{)} \\
= &2\left(\int_{\mathcal{X}} \hat{p}'(\boldsymbol{x}, y) - \hat{p}_{X|Y}(\boldsymbol{x}|y) + \hat{p}_{X|Y}(\boldsymbol{x}|y) + p^*_{X|Y}(\boldsymbol{x}|y) d\boldsymbol{x}\right) \\
&\left(\int_{\mathcal{X}} \hat{p}'(\boldsymbol{x}, y) - \hat{p}_{X|Y}(\boldsymbol{x}|y) + \hat{p}_{X|Y}(\boldsymbol{x}|y) + p^*_{X|Y}(\boldsymbol{x}|y) - 2\sqrt{\hat{p}'(\boldsymbol{x}, y) p^*_{X|Y}(\boldsymbol{x}|y)} d\boldsymbol{x}\right) \\
\leq &2(\epsilon+2)\left(\epsilon + 2 - 2\int_{\mathcal{X}} \sqrt{\hat{p}'(\boldsymbol{x}, y) p^*_{X|Y}(\boldsymbol{x}|y)} d\boldsymbol{x}\right). \\
&\qquad \text{(by } \textstyle\int_{\mathcal{X}} |\hat{p}_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x} = \int_{\mathcal{X}} |p^*_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x} = 1 \text{ and } \int_{\mathcal{X}} |\hat{p}'(\boldsymbol{x}, y) - \hat{p}_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x} \leq \epsilon\text{)}
\end{aligned}
$$

Putting together (I) and (II), we get:

$$
\begin{aligned}
\mathrm{TV}(\hat{p}_{X|Y}, p^*_{X|Y}) &= \frac{1}{2}\sqrt{\left(\int_{\mathcal{X}} |\hat{p}_{X|Y}(\boldsymbol{x}|y) - p^*_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x}\right)^2} \\
&\leq \frac{1}{2}\sqrt{(\epsilon+4)\epsilon + 2(\epsilon+2)\left(\epsilon + 2 - 2\int_{\mathcal{X}} \sqrt{\hat{p}'(\boldsymbol{x}, y) p^*_{X|Y}(\boldsymbol{x}|y)} d\boldsymbol{x}\right)}. \qquad (10)
\end{aligned}
$$

Bounding the average TV error. Based on the above results, we upper bound the average TV error (defined in Equation (5)) of $\hat{p}_{X|Y}$ as follows:

$$
\begin{aligned}
&\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}_{X|Y}) \\
= &\mathbb{E}_Y\Big[\mathrm{TV}(\hat{p}_{X|Y}, p^*_{X|Y})\Big] \\
\leq &\frac{1}{2}\mathbb{E}_Y\left[\sqrt{(\epsilon+4)\epsilon + 2(\epsilon+2)\left(\epsilon + 2 - 2\int_{\mathcal{X}} \sqrt{\hat{p}'(\boldsymbol{x}, y) p^*_{X|Y}(\boldsymbol{x}|y)} d\boldsymbol{x}\right)}\right] \qquad \text{(by Equation (10))} \\
\leq &\frac{1}{2}\sqrt{\mathbb{E}_Y\left[(\epsilon+4)\epsilon + 2(\epsilon+2)\left(\epsilon + 2 - 2\int_{\mathcal{X}} \sqrt{\hat{p}'(\boldsymbol{x}, y) p^*_{X|Y}(\boldsymbol{x}|y)} d\boldsymbol{x}\right)\right]} \\
&\qquad \text{(by concavity of } f(x) = \sqrt{x} \text{ and } \textit{Jensen's inequality}\text{)} \\
= &\frac{1}{2}\sqrt{(\epsilon+4)\epsilon + 2(\epsilon+2)\left(\epsilon + 2\left(1 - \mathbb{E}_Y\left[\int_{\mathcal{X}} \sqrt{\hat{p}'(\boldsymbol{x}, y) p^*_{X|Y}(\boldsymbol{x}|y)} d\boldsymbol{x}\right]\right)\right)}. \\
&\qquad \text{(by the linearity of expectation)}
\end{aligned}
$$

Recalling the elementary inequality we derived formerly in Equation (9), we have: it holds with at least probability $1-\delta$ that

$$\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}_{X|Y}) \leq \frac{1}{2}\sqrt{(\epsilon+4)\epsilon + 2(\epsilon+2)\left(\epsilon + \frac{2}{n}\log\frac{\mathcal{N}_{[]}\left(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta}\right)}. \tag{11}$$

Recalling that $0 \leq \delta \leq \frac{1}{2}$ and for non-empty $\mathcal{P}_{X|Y}$, $\mathcal{N}_{[]}\left(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right) \geq 1$, we have $\mathcal{N}_{[]}\left(\epsilon; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)/\delta \geq 2 \geq e^{\frac{1}{2}}$. Taking $\epsilon = 1/n$ in Equation (11), it then holds with probability at least $1-\delta$ that

$$\begin{aligned}
\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}_{X|Y}) &\leq \frac{1}{2}\sqrt{(\frac{1}{n}+4)\frac{1}{n} + 2(\frac{1}{n}+2)\left(\frac{1}{n} + \frac{2}{n}\log\frac{\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta}\right)} \\
&\leq \frac{1}{2}\sqrt{\frac{5}{n} + 6\left(\frac{1}{n} + \frac{2}{n}\log\frac{\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta}\right)} \qquad (\text{by } \tfrac{1}{n} \leq 1) \\
&\leq \frac{1}{2}\sqrt{\frac{10}{n}\log\frac{\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta} + 6\left(\frac{4}{n}\log\frac{\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta}\right)} \\
&\qquad\qquad (\text{by } \tfrac{1}{n} \leq \tfrac{2}{n}\log\tfrac{\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta}) \\
&= \frac{1}{2}\sqrt{\frac{34}{n}\log\frac{\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta}} \leq 3\sqrt{\frac{1}{n}\log\frac{\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right)}{\delta}} \\
&= 3\sqrt{\frac{1}{n}\left(\log\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}_{X|Y}, L^1(\mathcal{X})\right) + \log\frac{1}{\delta}\right)}.
\end{aligned} \tag{12}$$

Until now, we have completed the proof of this theorem.

□

B.3 Proof of Proposition 3.2

Proof of Proposition 3.2. As defined in Section 2, it holds that $\mathcal{P}^{\mathrm{multi}}_{X|Y} \subset \mathcal{P}^{\mathrm{single}}_{X|Y}$. Then, for any $p^{\mathrm{multi}}_{X|Y} \in \mathcal{P}^{\mathrm{multi}}_{X|Y}$, there exists some $p^{\mathrm{single}}_{X|Y} \in \mathcal{P}^{\mathrm{single}}_{X|Y}$ such that $p^{\mathrm{single}}_{X|Y} = p^{\mathrm{multi}}_{X|Y}$. Given any $\epsilon > 0$ and $1 \leq \mathsf{p} \leq \infty$, let $\mathcal{B}^{\mathrm{single}}$ be a ϵ-upper bracket w.r.t. $L^{\mathsf{p}}(\mathcal{X})$ for $\mathcal{P}^{\mathrm{single}}_{X|Y}$ such that $|\mathcal{B}^{\mathrm{single}}| = \mathcal{N}_{[]}\left(\epsilon; \mathcal{P}^{\mathrm{single}}_{X|Y}, L^{\mathsf{p}}(\mathcal{X})\right)$. According to the definition of ϵ-upper bracket (as in Definition B.1), there exists some $p' \in \mathcal{B}^{\mathrm{single}}$ such that given any $y \in \mathcal{Y}$, it holds that: $\forall \boldsymbol{x} \in \mathcal{X}, p'(\boldsymbol{x}, y) \geq p^{\mathrm{single}}_{X|Y}(\boldsymbol{x}|y) = p^{\mathrm{multi}}_{X|Y}(\boldsymbol{x}|y)$, and $\|p'(\cdot, y) - p^{\mathrm{multi}}_{X|Y}(\cdot|y)\|_{L^{\mathsf{p}}(\mathcal{X})} = \|p'(\cdot, y) - p^{\mathrm{single}}_{X|Y}(\cdot|y)\|_{L^{\mathsf{p}}(\mathcal{X})} \leq \epsilon$. Therefore, $\mathcal{B}^{\mathrm{single}}$ is also a ϵ-upper bracket w.r.t. $L^{\mathsf{p}}(\mathcal{X})$ for $\mathcal{P}^{\mathrm{multi}}_{X|Y}$, and thus $\mathcal{N}_{[]}\left(\epsilon; \mathcal{P}^{\mathrm{multi}}_{X|Y}, L^{\mathsf{p}}(\mathcal{X})\right) \leq |\mathcal{B}^{\mathrm{single}}| = \mathcal{N}_{[]}\left(\epsilon; \mathcal{P}^{\mathrm{single}}_{X|Y}, L^{\mathsf{p}}(\mathcal{X})\right)$. □

C PROOFS FOR SECTION 4

C.1 BRACKETING NUMBER OF CONDITIONAL GAUSSIAN DISTRIBUTION SPACE

According to Theorem 3.1, to derive the upper bound of average TV error, we need to measure the upper bracketing number for the conditional Gaussian distribution space. This result mainly follows the bracketing number analysis of Gaussian distribution space in Lemma C.5 in (Ge et al., 2024) and slightly modifies it to conditional Gaussian distribution space.

Theorem C.1 (Bracketing number upper bound for conditional Gaussian distribution space under multi-source training)**.** *Let B be a constant that $0 < B < \infty$, suppose that $\Phi = [-B, B]^{d_1}$, $\Psi = [-B, B]^{d-d_1}$, and conditional distributions in $\mathcal{P}_{X|Y}^{\text{multi}}$ are formulated as in Equation* (6)*. Then, given any $0 < \epsilon \leq 1$, the ϵ-upper bracketing number of $\mathcal{P}_{X|Y}^{\text{multi}}$ w.r.t. $L^1(\mathcal{X})$ satisfies*

$$\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}_{X|Y}^{\text{multi}}, L^1(\mathcal{X})\Big) \leq \left(\frac{2(1+d)B}{\epsilon} + 1\right)^{(K-1)d_1+d}.$$

Proof. According to the assumptions, the conditional distribution space expressed by the parametric estimation model is

$$\begin{aligned}\mathcal{P}_{X|Y}^{\text{multi}} \coloneqq \Bigg\{ & p_{X|Y}^{\text{multi}}(\boldsymbol{x}|y) = \prod_{k=1}^{K} \big(p_{\phi_k,\psi}(\boldsymbol{x}|k)\big)^{\mathbb{I}(y=k)} \\ & = \prod_{k=1}^{K} \Big((2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|\boldsymbol{x}-(\phi_k,\psi)\|_2^2}\Big)^{\mathbb{I}(y=k)} : \phi_k \in [-B,B]^{d_1}, \psi \in [-B,B]^{d-d_1} \Bigg\}.\end{aligned}$$

For any $p_{X|Y}^{\text{multi}}(\boldsymbol{x}|y) = \prod_{k=1}^{K} \Big((2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|\boldsymbol{x}-(\phi_k,\psi)\|_2^2}\Big)^{\mathbb{I}(y=k)} \in \mathcal{P}_{X|Y}^{\text{multi}}$, let's first divide the mean vector (ϕ_k, ψ) into η-width grids with a small constant $\eta > 0$ (the value of η will be specified later): If $(\phi_k)_i \in [j\eta, (j+1)\eta)$ for some $j \in \mathbb{Z}$, let $(\bar{\phi}_k)_i = j\eta$ and $\bar{\phi}_k \coloneqq \big((\bar{\phi}_k)_1, \ldots, (\bar{\phi}_k)_{d_1}\big)$. Similarly, if $(\psi)_i \in [j\eta, (j+1)\eta)$ for some $j \in \mathbb{Z}$, let $(\bar{\psi})_i = j\eta$ and $\bar{\psi} \coloneqq \big((\bar{\psi})_1, \ldots, (\bar{\psi})_{d-d_1}\big)$. In this case, we have $\|(\phi_k, \psi) - (\bar{\phi}_k, \bar{\psi})\|_2^2 \leq d\eta^2$.

Let

$$p'(\boldsymbol{x}, y) = \prod_{k=1}^{K} \Big((2\pi)^{-\frac{d}{2}} e^{-\frac{c_1}{2}\|\boldsymbol{x}-(\bar{\phi}_k,\bar{\psi})\|_2^2 + c_2}\Big)^{\mathbb{I}(y=k)}.$$

According to the definition of the bracketing, we want to prove that $p'(\boldsymbol{x}, y) \geq p_{X|Y}^{\text{multi}}(\boldsymbol{x}|y)$. By completing the square w.r.t. $\boldsymbol{x}$, we have

$$\begin{aligned} & -\frac{c_1}{2}\|\boldsymbol{x} - (\bar{\phi}_k, \bar{\psi})\|_2^2 + c_2 - \left(-\frac{1}{2}\|\boldsymbol{x} - (\phi_k, \psi)\|_2^2\right) \\ = & \frac{1}{2}\left((1-c_1)\left\|\boldsymbol{x} + \frac{c_1(\bar{\phi}_k, \bar{\psi}) - (\phi_k, \psi)}{1-c_1}\right\|_2^2 - \frac{c_1}{1-c_1}\big\|(\bar{\phi}_k, \bar{\psi}) - (\phi_k, \psi)\big\|_2^2 + 2c_2\right).\end{aligned}$$

Further taking $c_1 = 1 - \eta$ and $c_2 = d(1-\eta)\eta/2$, we have

$$\begin{aligned} & (1-c_1)\left\|\boldsymbol{x} + \frac{c_1(\bar{\phi}_k, \bar{\psi}) - (\phi_k, \psi)}{1-c_1}\right\|_2^2 - \frac{c_1}{1-c_1}\big\|(\bar{\phi}_k, \bar{\psi}) - (\phi_k, \psi)\big\|_2^2 + 2c_2 \\ = & \eta\left\|\boldsymbol{x} + \frac{c_1(\bar{\phi}_k, \bar{\psi}) - (\phi_k, \psi)}{1-c_1}\right\|_2^2 - \frac{1-\eta}{\eta}\big\|(\bar{\phi}_k, \bar{\psi}) - (\phi_k, \psi)\big\|_2^2 + 2c_2 && (c_1 = 1-\eta) \\ \geq & -\frac{1-\eta}{\eta}\big\|(\bar{\phi}_k, \bar{\psi}) - (\phi_k, \psi)\big\|_2^2 + 2c_2 && (\eta > 0) \\ \geq & -\frac{1-\eta}{\eta} d\eta^2 + 2c_2 && (\|(\phi_k, \psi) - (\bar{\phi}_k, \bar{\psi})\|_2^2 \leq d\eta^2) \\ = & -d(1-\eta)\eta + d(1-\eta)\eta = 0.\end{aligned}$$

Therefore, it holds that for all $y \in \mathcal{Y}$,

$$\forall \boldsymbol{x} \in \mathcal{X} : p'(\boldsymbol{x}, y) \geq p^{\text{multi}}_{X|Y}(\boldsymbol{x}|y). \tag{13}$$

Moreover, given any $0 < \epsilon \leq 1$, we take $\eta = \frac{\epsilon}{1+d}$, and thus $c_1 = 1 - \frac{\epsilon}{1+d}$ and $c_2 = \frac{1}{2}(1 - \frac{\epsilon}{1+d})\frac{\epsilon}{\frac{1}{d}+1}$. Since $d \in \mathbb{N}$, we have $\eta \leq \frac{1}{2}$ and $c_2 \leq \frac{1}{2}$. Then, $\|p'(\cdot, y) - p^{\text{multi}}_{X|Y}(\cdot|y)\|_{L^1(\mathcal{X})}$ can be bounded as

$$\begin{aligned}
&\|p'(\cdot, y) - p^{\text{multi}}_{X|Y}(\cdot|y)\|_{L^1(\mathcal{X})} = \int_{\mathcal{X}} |p'(\boldsymbol{x}, y) - p^{\text{multi}}_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x} \\
=& \int_{\mathcal{X}} p'(\boldsymbol{x}, y) d\boldsymbol{x} - \int_{\mathcal{X}} p^{\text{multi}}_{X|Y}(\boldsymbol{x}|y) d\boldsymbol{x} = \frac{1}{\sqrt{c_1}} e^{c_2} - 1 && (\textstyle\int_{\mathcal{X}} e^{-\frac{1}{2}\|\boldsymbol{x}\|_2^2} d\boldsymbol{x} = (2\pi)^{\frac{d}{2}}) \\
\leq& \frac{1}{\sqrt{c_1}}(1 + 2c_2) - 1 && (e^x \leq 1 + 2x \text{ for } x \in [0, \tfrac{1}{2}]) \\
=& \frac{1}{\sqrt{1-\eta}}(1 + d(1-\eta)\eta) - 1 && (c_1 = 1 - \eta \text{ and } c_2 = d(1-\eta)\eta/2) \\
\leq& (1 + \eta)(1 + d(1-\eta)\eta) - 1 && (\tfrac{1}{\sqrt{1-x}} \leq 1 + x \text{ for } x \in [0, \tfrac{1}{2}]) \\
=& \eta\Big(1 + d(1 - \eta^2)\Big) \leq \eta(1 + d) = \epsilon
\end{aligned} \tag{14}$$

Combining Equation (13) and Equation (14), we know that for any $p^{\text{multi}}_{X|Y}(\boldsymbol{x}|y) \in \mathcal{P}^{\text{multi}}_{X|Y}$ and $0 < \epsilon \leq 1$, there exists some $p'(\boldsymbol{x}, y) \in \mathcal{B}$ such that given any $y \in \mathcal{Y}$, it holds that $\forall \boldsymbol{x} \in \mathcal{X} : p'(\boldsymbol{x}, y) \geq p_{X|Y}(\boldsymbol{x}|y)$, and $\|p'(\cdot, y) - p_{X|Y}(\cdot|y)\|_{L^p(\mathcal{X})} \leq \epsilon$, where

$$\mathcal{B} \coloneqq \left\{ p'(\boldsymbol{x}, y) = \prod_{k=1}^{K} \Big((2\pi)^{-\frac{d}{2}} e^{-\frac{c_1}{2}\|\boldsymbol{x} - (\bar{\phi}_k, \bar{\psi})\|_2^2 + c_2} \Big)^{\mathbb{I}(y=k)} : (\bar{\phi}_k)_i, (\bar{\psi})_i \in [-B, B] \cap \eta\mathbb{Z} \right\}$$

Recalling the definition of the upper bracketing number in Definition B.1, we know that $\mathcal{B}$ is an ϵ-upper bracket of $\mathcal{P}^{\text{multi}}_{X|Y}$ w.r.t. $L^1(\mathcal{X})$. Therefore,

$$\begin{aligned}
&\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}^{\text{multi}}_{X|Y}, L^1(\mathcal{X})\Big) \\
\leq& |\mathcal{B}| = \left| \Big\{ \{\bar{\phi}_k\}_{k=1}^K, \bar{\psi} : (\bar{\phi}_k)_i, (\bar{\psi})_i \in [-B, B] \cap \eta\mathbb{Z} \Big\} \right| \\
\leq& \left(\frac{2B}{\eta} + 1 \right)^{Kd_1 + d - d_1} \\
=& \left(\frac{2(1+d)B}{\epsilon} + 1 \right)^{(K-1)d_1 + d},
\end{aligned}$$

which completes the proof. □

C.2 Proof of Theorem 4.1

Proof of Theorem 4.1. As $\phi_k^* \in \Phi, \psi^* \in \Psi$, and $\hat{p}^{\text{multi}}_{X|Y}$ is the maximizer of likelihood $L_S(p_{X|Y})$ in $\mathcal{P}^{\text{multi}}_{X|Y}$, according to Theorem 3.1, we know that

$$\mathcal{R}_{\overline{\text{TV}}}(\hat{p}^{\text{multi}}_{X|Y}) \leq 3\sqrt{\frac{1}{n}\left(\log \mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}^{\text{multi}}_{X|Y}, L^1(\mathcal{X}) \right) + \log \frac{1}{\delta} \right)}.$$

According to Theorem C.1, it holds that

$$\mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}^{\text{multi}}_{X|Y}, L^1(\mathcal{X}) \right) \leq \big(2(1+d)Bn + 1 \big)^{(K-1)d_1 + d}.$$

Therefore, we obtain the result that

$$\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{multi}}_{X|Y}) \leq 3\sqrt{\frac{1}{n}\bigg(\big((K-1)d_1+d\big)\log\big(2(1+d)Bn+1\big)+\log\frac{1}{\delta}\bigg)}.$$

Omitting constants about n, K, d_1, d, B, and the logarithm term we have $\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{multi}}_{X|Y}) = \tilde{\mathcal{O}}\bigg(\sqrt{\frac{(K-1)d_1+d}{n}}\bigg)$. □

C.3 Average TV error bound under single-source training

Theorem C.2 (Average TV error bound for conditional Gaussian distribution space under single–source training)**.** *Let $\hat{p}^{\mathrm{single}}_{X|Y}$ be the likelihood maximizer defined in Equation* (4) *given $\mathcal{P}^{\mathrm{single}}_{X|Y}$ with conditional distributions as in Equation* (6)*. Suppose $\Phi = [-B,B]^{d_1}$, $\Psi = [-B,B]^{d-d_1}$ with constant $B>0$, and $\phi^*_k \in \Phi$, $\psi^* \in \Psi$. Then, for any $0<\delta\leq 1/2$, it holds with probability at least $1-\delta$ that*

$$\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{single}}_{X|Y}) = \tilde{\mathcal{O}}\bigg(\sqrt{\frac{Kd}{n}}\bigg).$$

Proof. The proof is very similar to that in the multi-source case. According to the assumptions, the conditional distribution space expressed by the parametric estimation model is $\mathcal{P}^{\mathrm{single}}_{X|Y} := \Big\{p^{\mathrm{single}}_{X|Y}(\boldsymbol{x}|y) = \prod_{k=1}^K \big\{p_{\phi_k,\psi_k}(\boldsymbol{x}|k)\big\}^{\mathbb{I}(y=k)} = \prod_{k=1}^K \Big((2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|\boldsymbol{x}-(\phi_k,\psi_k)\|_2^2}\Big)^{\mathbb{I}(y=k)} : \phi_k \in [-B,B]^{d_1}, \psi_k \in [-B,B]^{d-d_1}\Big\}$. For any $p^{\mathrm{single}}_{X|Y}(\boldsymbol{x}|y) = \prod_{k=1}^K \Big((2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|\boldsymbol{x}-(\phi_k,\psi_k)\|_2^2}\Big)^{\mathbb{I}(y=k)} \in \mathcal{P}^{\mathrm{single}}_{X|Y}$, let's first divide the mean vector (ϕ_k,ψ_k) into η-width grids with a small constant $\eta>0$ (the value of η will be specified later): If $(\phi_k)_i \in [j\eta,(j+1)\eta)$ for some $j\in\mathbb{Z}$, let $(\bar{\phi}_k)_i = j\eta$ and $\bar{\phi}_k := \big((\bar{\phi}_k)_1,\ldots,(\bar{\phi}_k)_{d_1}\big)$. Similarly, if $(\psi_k)_i \in [j\eta,(j+1)\eta)$ for some $j\in\mathbb{Z}$, let $(\bar{\psi}_k)_i = j\eta$ and $\bar{\psi}_k := \big((\bar{\psi}_k)_1,\ldots,(\bar{\psi}_k)_{d-d_1}\big)$. In this case, we have $\|(\phi_k,\psi_k)-(\bar{\phi}_k,\bar{\psi}_k)\|_2^2 \leq d\eta^2$.

Let

$$p'(\boldsymbol{x},y) = \prod_{k=1}^K \Big((2\pi)^{-\frac{d}{2}} e^{-\frac{c_1}{2}\|\boldsymbol{x}-(\bar{\phi}_k,\bar{\psi}_k)\|_2^2+c_2}\Big)^{\mathbb{I}(y=k)}.$$

We need $p'(\boldsymbol{x},y) \geq p^{\mathrm{single}}_{X|Y}(\boldsymbol{x}|y)$ by the definition of the bracketing. By completing the square w.r.t. $\boldsymbol{x}$, we have

$$\begin{aligned}
&-\frac{c_1}{2}\|\boldsymbol{x}-(\bar{\phi}_k,\bar{\psi}_k)\|_2^2+c_2-\bigg(-\frac{1}{2}\|\boldsymbol{x}-(\phi_k,\psi_k)\|_2^2\bigg)\\
=&\frac{1}{2}\bigg((1-c_1)\bigg\|\boldsymbol{x}+\frac{c_1(\bar{\phi}_k,\bar{\psi}_k)-(\phi_k,\psi_k)}{1-c_1}\bigg\|_2^2-\frac{c_1}{1-c_1}\big\|(\bar{\phi}_k,\bar{\psi}_k)-(\phi_k,\psi_k)\big\|_2^2+2c_2\bigg).
\end{aligned}$$

Further taking $c_1 = 1-\eta$ and $c_2 = d(1-\eta)\eta/2$, we have

$$\begin{aligned}
&(1-c_1)\bigg\|\boldsymbol{x}+\frac{c_1(\bar{\phi}_k,\bar{\psi}_k)-(\phi_k,\psi_k)}{1-c_1}\bigg\|_2^2-\frac{c_1}{1-c_1}\big\|(\bar{\phi}_k,\bar{\psi}_k)-(\phi_k,\psi_k)\big\|_2^2+2c_2\\
=&\eta\bigg\|\boldsymbol{x}+\frac{c_1(\bar{\phi}_k,\bar{\psi}_k)-(\phi_k,\psi_k)}{1-c_1}\bigg\|_2^2-\frac{1-\eta}{\eta}\big\|(\bar{\phi}_k,\bar{\psi}_k)-(\phi_k,\psi_k)\big\|_2^2+2c_2 && (c_1=1-\eta)\\
\geq&-\frac{1-\eta}{\eta}\big\|(\bar{\phi}_k,\bar{\psi}_k)-(\phi_k,\psi_k)\big\|_2^2+2c_2 && (\eta>0)\\
\geq&-\frac{1-\eta}{\eta}d\eta^2+2c_2 && (\|(\phi_k,\psi)-(\bar{\phi}_k,\bar{\psi}_k)\|_2^2\leq d\eta^2)\\
=&-d(1-\eta)\eta+d(1-\eta)\eta=0.
\end{aligned}$$

Therefore, it holds that for all $y \in \mathcal{Y}$,

$$\forall \boldsymbol{x} \in \mathcal{X} : p'(\boldsymbol{x}, y) \geq p^{\text{single}}_{X|Y}(\boldsymbol{x}|y). \tag{15}$$

Moreover, given any $0 < \epsilon \leq 1$, we take $\eta = \frac{\epsilon}{1+d}$, and thus $c_1 = 1 - \frac{\epsilon}{1+d}$ and $c_2 = \frac{1}{2}(1 - \frac{\epsilon}{1+d})\frac{\epsilon}{\frac{1}{d}+1}$. Since $d \in \mathbb{N}$, we have $\eta \leq \frac{1}{2}$ and $c_2 \leq \frac{1}{2}$. Then, $\|p'(\cdot, y) - p^{\text{single}}_{X|Y}(\cdot|y)\|_{L^1(\mathcal{X})}$ can be bounded as

$$\begin{aligned}
&\|p'(\cdot, y) - p^{\text{single}}_{X|Y}(\cdot|y)\|_{L^1(\mathcal{X})} = \int_{\mathcal{X}} |p'(\boldsymbol{x}, y) - p^{\text{single}}_{X|Y}(\boldsymbol{x}|y)| d\boldsymbol{x} \\
=& \int_{\mathcal{X}} p'(\boldsymbol{x}, y) d\boldsymbol{x} - \int_{\mathcal{X}} p^{\text{single}}_{X|Y}(\boldsymbol{x}|y) d\boldsymbol{x} = \frac{1}{\sqrt{c_1}} e^{c_2} - 1 && (\textstyle\int_{\mathcal{X}} e^{-\frac{1}{2}\|\boldsymbol{x}\|_2^2} d\boldsymbol{x} = (2\pi)^{\frac{d}{2}}) \\
\leq& \frac{1}{\sqrt{c_1}}(1 + 2c_2) - 1 && (e^x \leq 1 + 2x \text{ for } x \in [0, \tfrac{1}{2}]) \\
=& \frac{1}{\sqrt{1-\eta}}(1 + d(1-\eta)\eta) - 1 && (c_1 = 1 - \eta \text{ and } c_2 = d(1-\eta)\eta/2) \\
\leq& (1 + \eta)(1 + d(1-\eta)\eta) - 1 && (\tfrac{1}{\sqrt{1-x}} \leq 1 + x \text{ for } x \in [0, \tfrac{1}{2}]) \\
=& \eta\Big(1 + d(1 - \eta^2)\Big) \leq \eta(1 + d) = \epsilon
\end{aligned} \tag{16}$$

Combining Equation (15) and Equation (16), we know that for any $p^{\text{single}}_{X|Y}(\boldsymbol{x}|y) \in \mathcal{P}^{\text{single}}_{X|Y}$ and $0 < \epsilon \leq 1$, there exists some $p'(\boldsymbol{x}, y) \in \mathcal{B}$ such that given any $y \in \mathcal{Y}$, it holds that $\forall \boldsymbol{x} \in \mathcal{X} : p'(\boldsymbol{x}, y) \geq p_{X|Y}(\boldsymbol{x}|y)$, and $\|p'(\cdot, y) - p_{X|Y}(\cdot|y)\|_{L^p(\mathcal{X})} \leq \epsilon$, where

$$\mathcal{B} \coloneqq \left\{ p'(\boldsymbol{x}, y) = \prod_{k=1}^{K} \Big((2\pi)^{-\frac{d}{2}} e^{-\frac{c_1}{2}\|\boldsymbol{x} - (\bar{\phi}_k, \bar{\psi}_k)\|_2^2 + c_2} \Big)^{\mathbb{I}(y=k)} : (\bar{\phi}_k)_i, (\bar{\psi}_k)_i \in [-B, B] \cap \eta\mathbb{Z} \right\}$$

Recalling the definition of the upper bracketing number in Definition B.1, we know that $\mathcal{B}$ is an ϵ-upper bracket of $\mathcal{P}^{\text{single}}_{X|Y}$ w.r.t. $L^1(\mathcal{X})$. Therefore,

$$\begin{aligned}
&\mathcal{N}_{[]}\Big(\epsilon; \mathcal{P}^{\text{single}}_{X|Y}, L^1(\mathcal{X})\Big) \\
\leq& |\mathcal{B}| = \left| \Big\{ \{\bar{\phi}_k\}_{k=1}^K, \{\bar{\psi}_k\}_{k=1}^K : (\bar{\phi}_k)_i, (\bar{\psi}_k)_i \in [-B, B] \cap \eta\mathbb{Z} \Big\} \right| \\
\leq& \left(\frac{2B}{\eta} + 1 \right)^{Kd_1 + K(d - d_1)} \\
=& \left(\frac{2(1+d)B}{\epsilon} + 1 \right)^{Kd}.
\end{aligned}$$

Besides, according to Theorem 3.1, we know that

$$\begin{aligned}
\mathcal{R}_{\overline{\text{TV}}}(\hat{p}^{\text{single}}_{X|Y}) &\leq 3\sqrt{\frac{1}{n}\left(\log \mathcal{N}_{[]}\left(\frac{1}{n}; \mathcal{P}^{\text{single}}_{X|Y}, L^1(\mathcal{X}) \right) + \log \frac{1}{\delta} \right)} \\
&\leq 3\sqrt{\frac{1}{n}\Big(Kd \log\big(2(1+d)Bn + 1\big) + \log \frac{1}{\delta} \Big)}.
\end{aligned}$$

Omitting constants about n, K, d_1, d, B, and the logarithm term we have $\mathcal{R}_{\overline{\text{TV}}}(\hat{p}^{\text{multi}}_{X|Y}) = \tilde{\mathcal{O}}\Big(\sqrt{\frac{Kd}{n}} \Big)$. □

D SUPPLEMENTARY FOR EXPERIMENTS

D.1 SIMULATIONS ON CONDITIONAL GAUSSIAN ESTIMATION

In this part, we aim to examine the tightness of the derived upper bound that $\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{multi}}_{X|Y}) = \tilde{\mathcal{O}}(\sqrt{\frac{(K-1)d_1+d}{n}})$ in Theorem 4.1 and $\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{single}}_{X|Y}) = \tilde{\mathcal{O}}(\sqrt{\frac{Kd}{n}})$ in Theorem C.2.

In all of our simulations, we fix the data dimension $d = 10$ and $p^*_Y(k) = 1/K$ all $k \in [K]$. K, n, and the similarity factor $\beta_{\mathrm{sim}} = \frac{1-d_1}{d} \in [0, 1]$ are key parameters. The dissimilar dimension $d_1 = d - \lfloor \beta_{\mathrm{sim}} d \rfloor$. We set the source-specific feature as $\phi_k = k\mathbf{1} \in \mathbb{R}^{d_1}$ and the shared feature as $\psi = \mathbf{0} \in \mathbb{R}^{d-d_1}$. Under the setting of Section 4, conditional MLE has analytical solution as

$$\hat{\phi}_k = \textstyle\sum_{y_i=k} \boldsymbol{x}_i[1:d_1]/n_k,\ \hat{\psi} = \sum_{i=1}^{n} \boldsymbol{x}_i[d_1+1:d]/n$$

for multi-source training and

$$\hat{\phi}_k = \textstyle\sum_{y_i=k} \boldsymbol{x}_i[1:d_1]/n_k,\ \hat{\psi}_k = \sum_{y_i=k} \boldsymbol{x}_i[d_1+1:d]/n_k$$

for single-source training.

For evaluation, we randomly sample $n^{\mathrm{test}} = 500$ data points according to the true joint distribution $p^*_{X,Y}$. Empirically, we approximate the true TV distance by using the Monte Carlo method based on the test set, which can be written formally as

$$\mathcal{R}_{\overline{\mathrm{TV}}}(\hat{p}_{X|Y}) \approx \frac{1}{2n^{\mathrm{test}}} \sum_{i=1}^{n^{\mathrm{test}}} \left| \frac{\hat{p}_{X|Y}(\boldsymbol{x}_i|y_i)}{p^*_{X|Y}(\boldsymbol{x}_i|y_i)} - 1 \right| = \mathcal{R}^{\mathrm{em}}_{\overline{\mathrm{TV}}}(\hat{p}_{X|Y}).$$

To eliminate the randomness, we average over 5 random runs for each simulation and report the mean results.

Order of the average TV error about K. We range the number of sources K in $[1, 3, 5, 10, 15]$ with fixed sample size $n = 500$ and similarity factor $\beta_{\mathrm{sim}} = 0.5$. We display the empirical average TV error for each K in Figure 1(a), with $\mathcal{R}^{\mathrm{em}}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{multi}}_{X|Y})$ colored in green and $\mathcal{R}^{\mathrm{em}}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{single}}_{X|Y})$ colored in orange. Ignoring the influence of constants, it shows a good alignment between empirical errors (in solid lines) and theoretical upper bounds (in dashed lines), both scaling as $\tilde{\mathcal{O}}(\sqrt{K})$.

Order of the average TV error about n. We range sample size n in $[100, 300, 500, 1000, 5000]$ with fixed number of sources $K = 5$ and similarity factor $\beta_{\mathrm{sim}} = 0.5$. We display the empirical error for each n in Figure 1(b), with $\mathcal{R}^{\mathrm{em}}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{multi}}_{X|Y})$ colored in green and $\mathcal{R}^{\mathrm{em}}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{single}}_{X|Y})$ colored in orange. Ignoring the influence of constants, it shows that the orders of empirical error about n match well with the theoretical upper bounds which scale as $\tilde{\mathcal{O}}(1/\sqrt{n})$.

Order of the average TV error about β_{sim}. We range similarity factor β_{sim} in $[0, 0.3, 0.5, 0.7, 1]$ with fixed sample size $n = 500$ and number of data sources $K = 5$. We display the empirical average TV error for each β_{sim} in Figure 1(c) to observe how similarity factor β_{sim} impacts the advantage of multi-source training. Concretely, as predicted by the theoretical bounds, the changing of β_{sim} will not influence the performance of single-source training but will decrease the error of multi-source training in the order of $\tilde{\mathcal{O}}(\sqrt{d_1}) = \tilde{\mathcal{O}}(\sqrt{1-\beta_{\mathrm{sim}}})$. The results show that the theoretical bounds predict the empirical performance well.

To sum up, our simulations verify the validity of our theoretical bounds in Section 4. Moreover, in all experiments, $\mathcal{R}^{\mathrm{em}}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{multi}}_{X|Y})$ is consistently smaller than $\mathcal{R}^{\mathrm{em}}_{\overline{\mathrm{TV}}}(\hat{p}^{\mathrm{single}}_{X|Y})$, supporting our results in Section 3.

D.2 REAL-WORLD EXPERIMENTS ON DIFFUSION MODELS

In this section, we conduct experiments on diffusion models to validate our theoretical findings in real-world scenarios from two aspects: (1) We empirically compare multi-source and single-source training

on conditional diffusion models and evaluate their performance to validate the guaranteed advantage of multi-source training against single-source training proved in Section 3. (2) We investigate the trend of this advantage about key factors—the number of sources and distribution similarity—as discussed in Section 4.

Experimental settings. We train class-conditional diffusion models following EDM2 (Karras et al., 2024) at 256×256 resolution on the selected classes from the ILSVRC2012 training set (Russakovsky et al., 2015), which is a subset of ImageNet (Deng et al., 2009) containing 1.28M natural images from 1000 classes, each annotated with an integer class label from 1 to 1000. In our experiments, we treat each class as a distinct data source. To control similarity among data sources, we manually design two levels of distribution similarity based on the semantic hierarchy of ImageNet (Deng et al., 2010; Bostock., 2018) as shown in Figure 3.

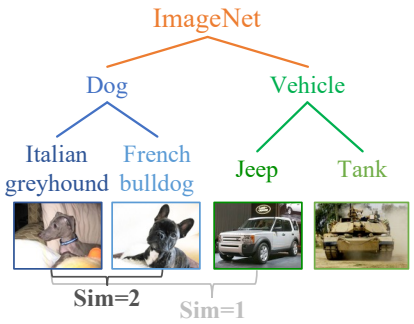

Figure 3: Similarity level.

Following EDM2, we use the Latent Diffusion Model (LDM) (Rombach et al., 2022) to down-sample each image $x \in \mathbb{R}^{3\times256\times256}$ to a corresponding latent $z \in \mathbb{R}^{4\times32\times32}$ for training a diffusion models. All experiments are trained and sampled on 8 × NVIDIA A800 80GB, 8 × NVIDIA GeForce RTX 4090, and 8 × NVIDIA GeForce RTX 3090 on the Linux Ubuntu-22.04 platform. For a fair comparison, we set different hyper-parameters for experiments with different numbers of sources as shown in Table 2, but these parameters are the same with different similarity levels.

Table 2: Hyparameters of our experiments. '1c' denotes training from single-source, and others denote training from multi-source which contains 3,5, and 10 classes.

Setup	Iterations (kimg)	Learning rate	Decay (kimg)
1c	184549	0.005	2500
3c	268435	0.006	4000
10c	1610612	0.012	6000

For each controlled experiment comparing multi-source and single-source training, we fix K target classes within one similarity level Sim and train the models on a dataset S consisting of N examples per class. Under multi-source training, we train a single conditional diffusion model for all K classes jointly. Under single-source training, we train K separate conditional diffusion models, one for each class. Please refer to Section 2 for the formal formulation of these two strategies. We set each factor with two possible values: the number of classes K in 3 or 10, distribution similarity Sim in 1 or 2, and the sample size per class N in 500 or 1000. This results in a total of 8 sets of experiments comparing multi-source and single-source training.

We evaluate model performance using the average Fréchet Inception Distance (Heusel et al., 2017) (FID, a widely used metric for image generation quality) across all conditions to assess the overall conditional generation performance. Results are displayed in Table 1. Specifically, for multi-source training, we compute the FID for each class and take the average over all K classes. For single-source training, we compute the FID for each of the K separately trained models on their respective

classes and calculate the average. Relative advantage of multi-source training is measured by $\frac{\text{Avg. FID (Single)} - \text{Avg. FID (Multi)}}{\text{Avg. FID (Single)}}$ as displayed in Figure 2.

Experimental results In the following, we interpret the results sequentially from the view of our theoretical findings.

From Table 1, we observe that under different amounts of classes K, similarity level Sim, and per-class sample size N, multi-source training generally achieves lower average FID than that of single-source training, which is consistent with our theoretical guarantees derived in Section 3,

From Figure 2, we observe that for any fixed similarity level Sim and per-class sample size N, the relative advantage of multi-sources training with a larger K (the green bars) is larger than that with a smaller K (the nearby orange bars). Additionally, for any fixed K and N, the relative advantage of multi-sources training with a larger distribution similarity is larger than that with a smaller distribution similarity (as shown through the dashed lines). These results support our theoretical insights in Section 4 that the number of sources and similarity among source distributions improves the advantage of multi-source training.