Controlling Mutual Information for Generative Few-Shot Data Augmentation via Parametric Distribution Construction

Zhiwei Sun, Zhuofan Chen, Jianfei Zhang, Wenge Rong, Zhang Xiong
School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{sunzhiwei, zhuofanchen, zhangjf, w.rong, xiongz}@buaa.edu.cn

***Abstract*—Data Augmentation (DA) remains a challenge for Natural Language Understanding in extreme low-resource scenarios. Although generative models can exploit auxiliary high-resource data to synthesize few-shot samples, current approaches typically oscillate between semantic drift and mode collapse. From an information-theoretic standpoint, we analyze the efficacy of generative DA through the lens of Mutual Information $I(X;Y)$ in the training distribution, and show that existing methods implicitly operate at rigid entropy extremes of this spectrum. To address this, we formulate *Parametric Distribution Construction*, a principled framework for regulating conditional entropy via tunable hyper-parameters, and instantiate it through two complementary strategies: Hard (cluster-based) and Soft (attention-based). Extensive evaluations across three real-world NLU benchmarks show that these instantiations effectively balance the correctness-diversity trade-off, filling the Pareto gap left by static baselines. They outperform the state-of-the-art PromDA, particularly on coarse-grained tasks characterized by high intra-class variance. Furthermore, entropy regulation serves as a robust safeguard against sampling bias, ensuring superior generalization even when seed data is non-representative.**

***Index Terms*—Data augmentation, Natural language understanding, Information theory, Mutual information.**

I. Introduction

Supervised NLU systems [1], [2] depend on large volumes of labeled data, and their accuracy degrades sharply when annotations are scarce. Data Augmentation (DA) alleviates this bottleneck by expanding training sets without additional annotation [3]. Augmentation methods mainly include word substitution [4]–[6], back translation [7]–[9], and feature perturbation [10]–[12]. Modern language models further provide a scalable generative paradigm to synthesize diverse, near-realistic text [13]–[16].

While these methods augment limited data through heuristic strategies, a major challenge lies in training NLU models to capture domain knowledge. Existing works exploit auxiliary in-domain high-resource (many-shot) data to learn end-to-end generative models for knowledge-enhanced data augmentation [17], [18], as shown in Fig. 1.

However, since many-shot data does not explicitly demonstrate how to generalize few-shot sentences, the construction of the training distribution is pivotal. Existing works oscillate between two extremes: (1) Reconstruction [19]–[21], acting as "soft oversampling" with high fidelity but limited variance; and (2) Paraphrasing [22], [23], introducing diversity but suffering from semantic drift. We formalize these as degenerate limits of an entropy spectrum, rather than independent design choices.

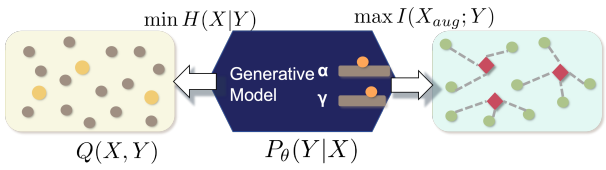


Fig. 1. The proposed GDA framework explicitly constructs a parametric training distribution $Q(X,Y)$ on in-domain many-shot data (left) to regulate mutual information, ensuring generalization to few-shot data (right).

We analyze generative effectiveness through the Mutual Information $I(X;Y)$ between source X and target Y within the training distribution. Under the assumption of a uniform label marginal $q(y)$ (which fixes $H(Y)$ and bounds $H(X)$ variation), we show that rigid distributions (high MI) lead to overfitting while chaotic ones (low MI) introduce noise. Parametric Distribution Construction is therefore essential for a tunable trade-off. This framework identifies *what* to control (conditional entropy), leaving *how* to control it as a secondary design choice.

As concrete instantiations, we propose Hard Parametric Construction (Cluster-based, parameter α) and Soft Parametric Construction (Attention-based, parameter γ), both quantitatively adjusting $I(X;Y)$ via hyper-parameters to avoid heuristic randomness.

Our contributions are three-fold: 1) We present a systematic information-theoretic characterization of GDA, formally establishing $I(X;Y)$ as the governing variable and deriving static baselines as degenerate entropy limits. 2) We demonstrate two parametric instantiations—Hard and Soft—that realize entropy regulation within the proposed framework, providing LLM-free, explicitly controllable augmentation. 3) Empirical results on three datasets validate the theoretical framework, with both instantiations filling the Pareto gap and outperforming PromDA [14]—demonstrating that explicit distribution control is more effective than scale alone.

II. Related Work

A. Generative Data Augmentation and LLMs

GDA has evolved from heuristic-rule baselines [4], [8] toward PLM-based synthesis [24], with LLMs—driven by ICL

and CoT—now redefining data generation for NLU [25]–[27]. Despite their potential, LLM-based augmentation is hampered by reliability and faithfulness issues: even SOTA models exhibit performance degeneration [28] and *hallucinations* [29] from uncontrollable internal entropy. In contrast, our work constructs an explicit, theoretically-grounded training distribution—characterized by its Mutual Information $I(X;Y)$—to regulate the generation of smaller, efficient models.

B. Information Theoretic Perspectives on Augmentation

Information Theory provides a rigorous lens to characterize the tension between semantic fidelity (Correctness) and lexical variance (Diversity). The Information Bottleneck (IB) principle [30] and its variational extensions [31] suggest optimal representations should balance compression with task-relevance. In DA, group-theoretic frameworks further emphasize capturing distributional invariance [32]. Our work specifically addresses the tension between maximizing mutual information $I(X_{aug};Y)$ for label consistency and maximizing entropy $H(X_{aug})$ for diversity. While previous methods often occupy static points on this curve [22], we propose a parametric framework to explicitly navigate the *Pareto frontier* through quantifiable control of mutual information. This positions $I(X;Y)$ not merely as a diagnostic metric but as an actionable design variable—one that can be directly manipulated through the construction of the training distribution $Q(X,Y)$.

III. Methodology

A. Problem Formulation

We define a task-oriented NLU environment with many-shot labels L_{many} (dataset S_{many}) and few-shot labels L_{few} (dataset S_{few}). We learn a generator $P_\theta(Y|X)$ on S_{many} to synthesize samples for L_{few}, constructing $S_{DA|few}$ by sampling $y \sim P_\theta(Y|X=x)$ for each $x \in S_{few}$, and train on $S_{few,DA} = S_{few} \cup S_{DA|few}$.

B. Training Distribution via Entropy Regulation

We formulate GDA as a distribution matching problem by optimizing $P_\theta(Y|X)$ to approximate a joint training distribution $Q(X,Y) = q(y)q(x|y)$ via MLE:

$$\hat{\theta} = \arg\max_{\theta} \mathbb{E}_{x,y\sim Q(X,Y)}[\log p_\theta(y|x)] \tag{1}$$

Crucially, while Eq. (1) is standard for conditional text generation, our contribution lies not in the generator architecture but in treating $q(x|y)$ as a *designable variable* rather than an implicit consequence of data collection. Under a uniform marginal $q(y)$, we analyze $q(x|y)$ through its Conditional Entropy $H(X|Y)_Q$, identifying existing heuristic methods as rigid instantiations at entropy limits:

(1) **Self-Reconstruction (Zero Entropy):** $q(x|y) = \delta_y(x)$ minimizes $H(X|Y)$ to 0. While maximizing $I(X;Y)$, it precipitates degenerate memorization and zero diversity;

(2) **Random Sampling (Maximum Entropy):** $q(x|y) = \mathcal{U}(S_{l_y})$ maximizes lexical diversity but overlooks intra-class semantic granularity, introducing noise into the supervision signal.

To bridge these extremes, we propose **Parametric Distribution Construction** $q(x|y;\lambda)$, where λ continuously navigates the entropy spectrum between these two degenerate limits.

C. Parametric Distribution Construction Strategies

To implement entropy regulation, we require a metric space in which $Sim(x,y)$ is semantically meaningful, so that $I(X;Y)$ under any $q(x|y;\lambda)$ can be computed and controlled. We fine-tune a BERT encoder [33] on S_{many} via InfoNCE loss, whose contrastive objective aligns geometric proximity with label consistency—directly enabling $I(X;Y)$ regulation across the entropy spectrum. Drawing on structural priors from metric-based few-shot learning—the discrete centroid clustering of Prototypical Networks [34] and the continuous attention weighting of Matching Networks [35]—we instantiate two complementary paradigms: a discrete structural approach (Hard) and a continuous probabilistic approach (Soft).

1) Hard Parametric Construction: Cluster-based: This paradigm partitions the semantic space S_l into $k_l = \lfloor |S_l|^\alpha \rfloor$ disjoint clusters, where $\alpha \in [0,1]$ is the control parameter. Assuming intra-cluster semantic invariance, we set $q(x|y) = 1/|C_{l_y,c_y}|$ for $x \in C_{l_y,c_y}$ and 0 otherwise, interpolating between Self-Reconstruction ($\alpha \to 1$) and Random Sampling ($\alpha \to 0$).

2) Soft Parametric Construction: Attention-based: We define $q(x|y)$ as a temperature-scaled Boltzmann distribution over all same-label utterances:

$$q(x|y;\gamma) = \frac{\exp(\gamma \cdot Sim(x,y))}{\sum_{x' \in S_{l_y}} \exp(\gamma \cdot Sim(x',y))} \tag{2}$$

where $\gamma \in [0,\infty)$ is the inverse temperature. $\gamma \to \infty$ recovers the Zero Entropy limit (Dirac measure), while $\gamma \to 0$ recovers the Maximum Entropy limit (Uniform distribution), enabling smooth regulation of $H(X|Y)$ for semantically relevant diversity.

IV. Experiments

This section empirically validates the theoretical framework established in Section III-B. We quantify the impact of entropy regulation via the proposed Hard and Soft instantiations across three key dimensions: Generalization, Correctness, and Diversity, and extend to ablation studies on α and γ to assess the framework's controllability.

A. Experimental Setup

Our generative backbone is built upon UniLM [36], first fine-tuned on S_{many} via a multi-task objective integrating MLM [33] and Seq2Seq reconstruction [37]. We fine-tune a `bert-base-uncased` [33] encoder on S_{many} via contrastive learning to establish the semantic metric space (Section III-C). Downstream NLU classifiers are similarly built upon `bert-base-uncased`. Implementation details are in Section IV-J.

TABLE I

GENERALIZATION ON CLINC-FULL AND CLINC-SMALL. PARAMETRIC INSTANTIATIONS VS. STATIC BASELINES AND PROMDA. SELF-RECON.: SELF-RECONSTRUCTION; RANDOM SAMP.: RANDOM SAMPLING.

Model	CLINC-FULL				CLINC-SMALL			
	1%-shot	2%-shot	5%-shot	10%-shot	1%-shot	2%-shot	5%-shot	10%-shot
Self-Recon.	76.5(1.2)	85.5(0.8)	94.6(0.6)	**96.9(0.3)**	48.3(2.2)	76.9(1.9)	**90.0(0.7)**	**94.5(0.2)**
Random Samp.	74.0(2.3)	69.8(0.9)	83.1(0.8)	86.6(2.1)	59.4(0.9)	70.0(2.1)	75.7(2.2)	84.5(2.5)
Recon-Tokens	71.8(2.1)	82.3(0.9)	93.2(0.4)	96.4(0.2)	49.7(2.1)	74.8(1.7)	87.9(1.3)	93.7(0.8)
Recon-Span	75.3(1.4)	82.3(0.9)	92.7(0.9)	96.2(0.3)	52.3(3.4)	73.6(1.9)	89.6(0.7)	93.4(0.4)
PromDA (T5-Large)	**88.1(0.0)**	**91.9(0.3)**	92.0(0.3)	92.4(0.6)	**87.1(1.1)**	**86.4(1.4)**	89.8(1.3)	91.8(0.3)
Hard Param. (α=0.7)	**85.2(0.4)**	**87.9(1.2)**	**94.7(0.5)**	96.3(0.5)	**63.6(2.8)**	**81.9(0.9)**	90.0(0.8)	94.5(0.4)
Soft Param. (γ=128)	81.8(1.7)	86.0(0.8)	94.1(0.4)	96.1(0.5)	59.4(1.6)	78.1(1.5)	88.4(1.1)	94.0(0.6)

TABLE II

GENERALIZATION ON UNIDA. PARAMETRIC INSTANTIATIONS OUTPERFORM LARGE-SCALE PROMDA ON THIS COARSE-GRAINED, LONG-TAILED DATASET, VALIDATING THAT EXPLICIT $I(X;Y)$ CONTROL IS MORE EFFECTIVE THAN PARAMETER SCALING UNDER HIGH INTRA-CLASS VARIANCE.

Model	UniDA			
	1%-shot	2%-shot	5%-shot	10%-shot
Self-Recon.	87.2(1.0)	89.7(0.8)	91.9(0.3)	**93.5(0.3)**
Random Samp.	79.6(0.6)	79.6(0.3)	79.4(0.7)	79.8(0.6)
Recon-Tokens	87.4(1.1)	90.2(0.3)	91.8(0.5)	**93.6(0.2)**
Recon-Span	87.8(0.3)	89.7(0.9)	92.1(0.5)	93.1(0.8)
PromDA (T5-Large)	79.5(0.5)	79.6(0.0)	79.7(0.3)	80.0(0.2)
Hard Param. (α=0.7)	**89.0(0.8)**	**90.9(0.3)**	**92.3(0.4)**	93.0(0.7)
Soft Param. (γ=128)	86.4(1.7)	89.1(0.7)	88.5(1.8)	88.7(2.2)

B. Datasets

We evaluate on three benchmarks covering distinct semantic granularities: CLINC-FULL [38] (150 balanced classes, 150 samples each); CLINC-SMALL [38] (a resource-constrained variant probing the framework under intensified scarcity, 100 samples per class); and UniDA [39], a coarse-grained, long-tailed dialogue act dataset (20 classes, 558K statements). We partition labels into L_{many} and L_{few}, constructing S_{few} by retaining 1%, 2%, 5%, and 10% of data via proportion-based sampling, progressively matching the full label scale for fair comparison. Reproducibility is ensured via systematic sampling after lexicographical sorting (Fig. 2a). Biased pools at $p \in \{1/2, 1/4\}$ probe robustness to non-representative seeds (Fig. 2b, c).

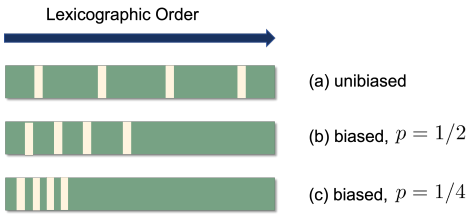


Fig. 2. Unbiased (a) and biased (b, c) few-shot sampling.

C. Baselines

The first two baselines empirically realize the theoretical entropy extremes of Section III-B, providing ground-truth anchors for the entropy spectrum.

Paraphrasing (Random Sampling) treats all instances as equiprobable inputs, representing the Maximum Entropy Limit:

$$q(x|y) = 1/|S_{l_y}|, \quad \text{for } x \in S_{l_y}. \tag{3}$$

Oversampling (Self-Reconstruction) enforces a strict identity mapping, serving as the Zero Entropy Limit:

$$q(x|y) = \mathbb{I}(x = y). \tag{4}$$

Reconstruction [19] applies Denoising Auto-Encoder masking (40% of tokens). **PromDA** [14] is a SOTA prompt-based method utilizing `T5-Large` (770M parameters), serving as a large-scale black-box reference point.

D. Metrics

Following [22], we evaluate $S_{few,DA}$ through three dimensions. **Generalization** trains a classifier on $S_{few,DA}$ and tests on $S_{few,true}$. **Correctness** applies reverse validation: training on $S_{few,true}$, testing on $S_{few,DA}$. **Diversity** uses Distinct-N [40] to measure unique n-gram ratios. Unlike distributional metrics requiring held-out real data, Correctness and Diversity directly characterize the intrinsic properties of $S_{few,DA}$, enabling evaluation independent of the true test distribution. We report mean and standard deviation across six runs (three for PromDA).

E. Main Results

Tables I–II validate the central claim: $I(X;Y)$ governs GDA efficacy. Parametric instantiations consistently outperform static baselines, with the most substantial gains at 1% shot. While static baselines are constrained to the derived entropy

Fig. 3. $I(X;Y)$ for different training distribution strategies and hyper-parameter settings on CLINC-FULL. Markers $+/\times$: static limits (Random/Self-Recon); $\diamond/\bullet$: Hard/Soft instantiations.

extremes, our instantiations navigate the full spectrum, with performance converging toward Self-Reconstruction at 10%—signaling diminishing returns as the underlying distribution densifies.

Second, our 110M-parameter framework demonstrates superior robustness on coarse-grained tasks over 770M-parameter PromDA: 89.0% vs. 79.5% on UniDA (1% shot). From an information-theoretic perspective, this failure stems from unconstrained black-box prompting's inability to regulate the conditional entropy of its implicit training distribution—resulting in uncontrolled $I(X;Y)$ that manifests as coverage failures under high intra-class variance. Our framework explicitly targets $I(X;Y)$ within the optimal entropy region, ensuring semantic coverage of coarse-grained labels' broad manifold. These results confirm that explicit distribution control is more effective than scale alone, precisely because it acts on the governing variable directly.

F. Ablation Study: Entropy Regulation

By varying $\alpha \in [0.1, 0.9]$ and $\gamma \in [0, 10^5]$, we regulate conditional entropy to observe the training distribution's behavior. As shown in Fig. 3, both parameters effectively control $I(X;Y)$. This direct quantification—ranging from 4.78 (Random Sampling) to 9.79 (Self-Reconstruction)—provides the first concrete measurement of an information-theoretic variable governing GDA, moving beyond qualitative notions of "diversity" or "faithfulness". Random Sampling anchors the lower bound (Maximum Entropy) and Self-Reconstruction the upper (Zero Entropy), with our instantiations spanning the full spectrum between semantic fidelity ($\alpha \to 1, \gamma \to \infty$) and lexical richness ($\alpha \to 0, \gamma \to 0$).

Fig. 4 consistently shows that Generalization follows a convex trajectory with respect to $I(X;Y)$: excessive noise at low MI degrades Correctness, while information scarcity at high MI suppresses Diversity. Peak Generalization is achieved at intermediate settings (Hard: $\alpha = 0.7$, Soft: $\gamma = 128$). Crucially, under 1% shot, Diversity drops sharply at the high-$I(X;Y)$ end while Correctness remains stable—explaining why our instantiations outperform Self-Reconstruction in low-resource regimes by injecting the diversity that small seed sets otherwise lack.

G. Ablation Study: Robustness to Sampling Bias

We vary $p \in \{1, 1/2, 1/4, 1/8\}$ (Fig. 2) to probe robustness against non-representative seeds, with results in Fig. 5. Lower-$I(X;Y)$ strategies—Random Sampling and Soft Parametric ($\gamma = 128$)—exhibit stronger resilience to bias, while high-MI strategies suffer significant degradation at $p = 1/8$ on UniDA. This follows directly from our framework: high-MI methods overfit the biased seed distribution, whereas lower-MI methods introduce stochastic variation that regularizes against narrow seeds. Formally, entropy acts as an implicit regularizer against sampling bias. Table III provides a qualitative illustration across α settings.

H. Case Study: Microscopic Analysis

Table III illustrates entropy regulation at the sample level. At low MI ($\alpha = 0.1$), maximized entropy maps inputs across 150 utterances, producing *semantic drift*. At high MI ($\alpha = 0.9$), near-deterministic mapping ($1/1$) causes *mode collapse*. The intermediate setting ($\alpha = 0.5$, $1/8$ neighborhood) achieves balanced paraphrase generation, confirming that entropy regulation directly controls the hallucination–memorization trade-off.

I. Discussion

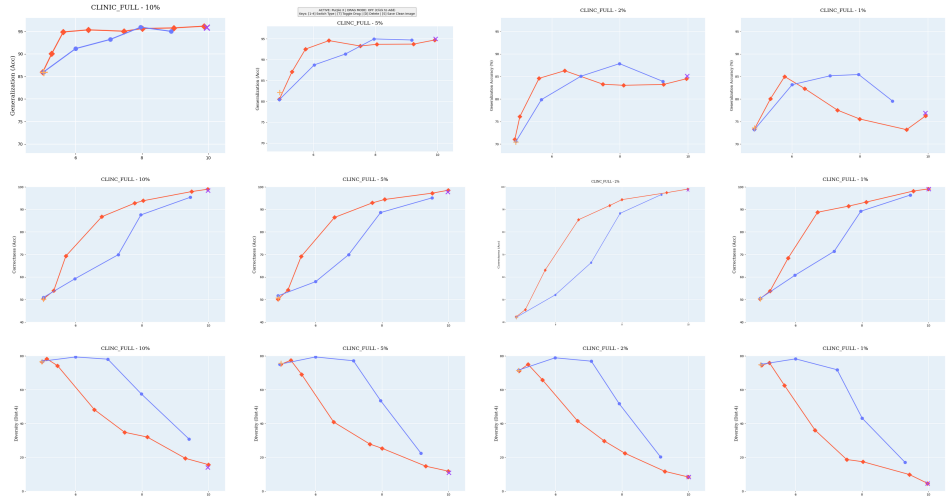
Our findings reveal that generative augmentation hinges on regulating information flow rather than data acquisition alone. While both instantiations traverse the entropy spectrum, their optimal operating points differ: the Hard instantiation's discrete structural bias enforces stricter semantic boundaries, while the Soft instantiation's continuous manifold assumption offers smoother but more calibration-sensitive transitions. A limitation is the reliance on a BERT-based metric optimized for discriminative tasks. Future work could explore alternative instantiation strategies within this framework—including top-k neighbor selection, LLM-based pattern curation, or generative-aware metric learning—each representing a different mechanism for the parametric control identified here.

J. Implementation Details

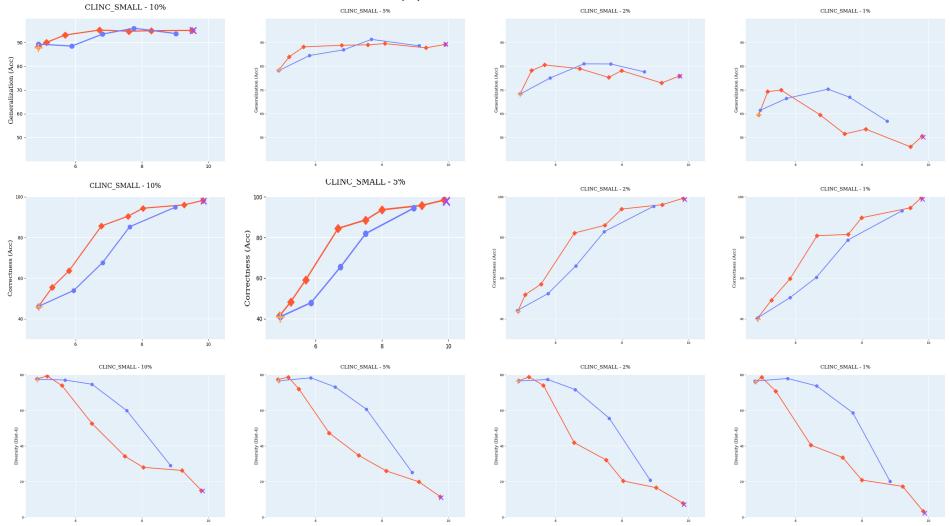
Our pipeline consists of three sequential phases—Semantic Metric Learning, Generative Pre-training, and Parametric Fine-tuning—as detailed in Table IV. We use `bert-base-uncased` [33] as the metric encoder and `unilm-base-cased` [36] as the generator, implemented with bert4keras [41] on a Tesla V100 GPU (16GB). Optimization uses AdamW [42] with batch size 32 across all phases.

V. Conclusion

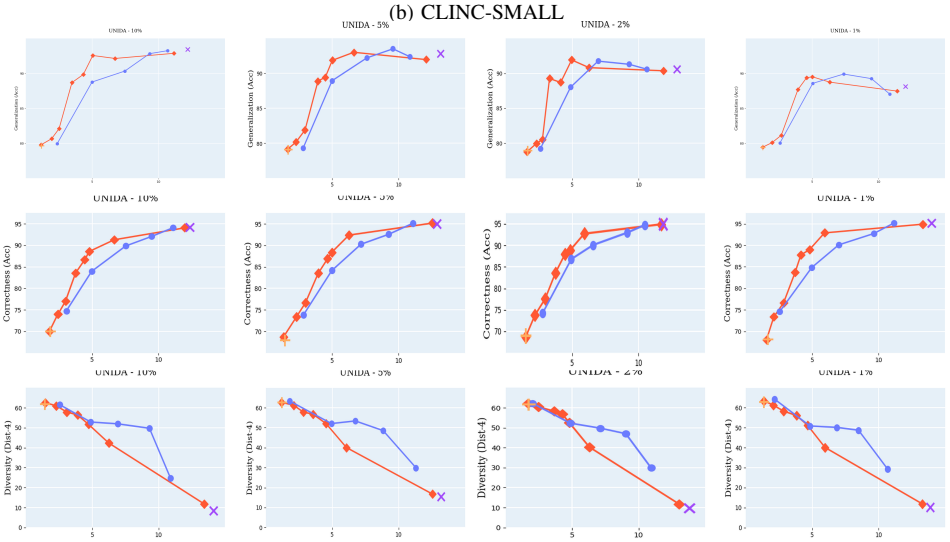
In this paper, we establish an information-theoretic framework for generative few-shot Natural Language Understanding. By formally identifying Mutual Information $I(X;Y)$ as the governing variable of generative augmentation efficacy and establishing static baselines as its degenerate entropy limits, we introduce Parametric Distribution Construction as the principled approach to regulate conditional entropy $H(X|Y)$, and demonstrate two concrete instantiations—Hard (cluster-based, α) and Soft (attention-based, γ). Extensive experiments



(a) CLINC-FULL



(b) CLINC-SMALL



(c) UniDA

Fig. 4. Generalization, Correctness, and Diversity of $S_{few,DA}$ vs. $I(X;Y)$ on CLINC-FULL (a), CLINC-SMALL (b), and UniDA (c).

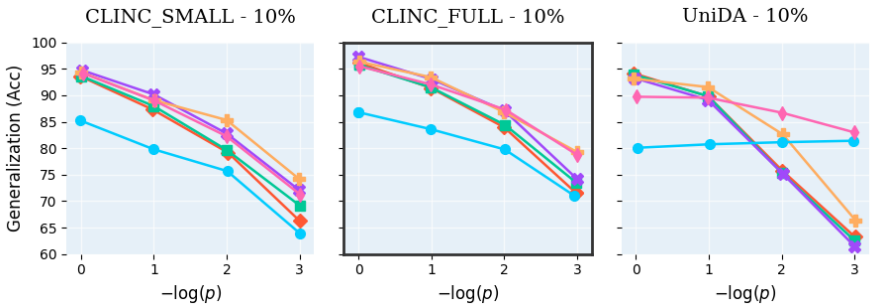


Fig. 5. Generalization under increasing bias $-\log(p)$. Markers: Recon-Span ($\Diamond$), Recon-Tokens ($\square$), Oversampling ($\times$), KMeans ($+$), Paraphrasing ($\bullet$), Attention ($*$).

TABLE III
CASE STUDY: ENTROPY REGULATION ON THE HARD PARAMETRIC INSTANTIATION (CLINC-FULL).

Many-shot Anchor (y_0)**:** "has the discover card approved my app"		
Few-shot Seed (x_0)**:** "i'm going to need a hotel in tampa near ybor city for march 23rd through the 28th"		
Hard Parametric ($\alpha = 0.1$) Low MI (4.78)	Training Dist. $q(x\|y_0)$	1/150: "has my application for the mastercard card gone through" 1/150: "when will my application for my credit card be processed"
	Generated Output $y \sim P_\theta(Y\|x_0)$	[False] "find me a place where i can rent a car" *(Semantic Drift)* [False] "i am going out of country and need a w2 form" [True] "is there a full friendly hotel in seattle"
Hard Parametric ($\alpha = 0.5$) Balanced MI (6.97)	Training Dist. $q(x\|y_0)$	1/8: "has the discover card approved my app" 1/8: "has visa gotten back to me regarding my card app"
	Generated Output $y \sim P_\theta(Y\|x_0)$	[True] "i need argentina destination from march 1-1" [True] "reserve a special place for new york food in seattle" [True] "use my house to book a visit to tourists in france"
Hard Parametric ($\alpha = 0.9$) High MI (9.05)	Training Dist. $q(x\|y_0)$	1/1: "has the discover card approved my app"
	Generated Output $y \sim P_\theta(Y\|x_0)$	[True] "i'm going to need a hotel in tampa near ybor city..." *(Repetition)* [True] "i'm going to need a hotel in tampa near ybor city..." [True] "i'm going to need a hotel in tampa near ybor city..."

TABLE IV
HYPERPARAMETERS, TRAINING STEPS, AND AVERAGE TIME COST.

Phase	Epochs	Learning Rate	Grad. Accum.	Steps per Epoch (Trn/Val)		Avg. Time Cost (seconds)		
				CLINC (S/F)	UniDA	SMALL	FULL	UniDA
Metric Learning	100	1e-5	1	100 / 50	1000 / 500	1276s	1273s	13323s
Gen. Pre-training	30	1e-4	16	100 / 50	1000 / 500	904s	2217s	8842s
Parametric Fine-tuning	100	1e-4	16	100 / 50	1000 / 500	2217s	2406s	29801s
NLU Gen. Evaluation	10	1e-4	16	375 / 93	50 / 12	172s	180s	111s
NLU Corr. Evaluation						138s	140s	38s

validate that both instantiations significantly outperform static baselines and achieve superior robustness over large-scale prompt-based models in low-resource scenarios. Our analysis reveals that augmentation effectiveness stems from calibrating $I(X;Y)$ between input and output, directly optimizing the correctness–diversity trade-off across the entropy spectrum. This theoretical insight is agnostic to the specific instantiation mechanism, providing a foundation for future extensions to alternative selection strategies and generation paradigms.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62477001.

REFERENCES

[1] P. Xu and R. Sarikaya, "Convolutional neural network based triangular CRF for joint intent detection and slot filling," in *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 78–83.

[2] Y. Chen, D. Hakkani-Tür, G. Tür, J. Gao, and L. Deng, "End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, 2016, pp. 3245–3249.

[3] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Computing Surveys*, vol. 56, no. 3, pp. 64:1–64:37, 2024.

[4] Q. Xie, Z. Dai, E. H. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020, pp. 6256–6268.

[5] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 452–457.

[6] J. W. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 6381–6387.

[7] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 567–573.

[8] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 86–96.

[9] A. W. Yu, D. Dohan, M. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "QANet: Combining local convolution with global self-attention for reading comprehension," in *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[10] D. Guo, Y. Kim, and A. M. Rush, "Sequence-level mixed sample data augmentation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 5547–5552.

[11] S. Zhang, X. Liu, J. Liu, J. Gao, K. Duh, and B. V. Durme, "ReCoRD: Bridging the gap between human and machine commonsense reading comprehension," *CoRR*, vol. abs/1810.12885, 2018.

[12] H. Yang and K. Li, "Boosting text augmentation via hybrid instance filtering framework," in *Findings of the Association for Computational Linguistics: ACL*, 2023, pp. 1652–1669.

[13] K. M. Yoo, D. Park, J. Kang, S. Lee, and W. Park, "GPT3Mix: Leveraging large-scale language models for text augmentation," in *Findings of the Association for Computational Linguistics: EMNLP*, 2021, pp. 2225–2239.

[14] Y. Wang, C. Xu, Q. Sun, H. Hu, C. Tao, X. Geng, and D. Jiang, "PromDA: Prompt-based data augmentation for low-resource NLU tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 4242–4255.

[15] J. Zhou, Y. Zheng, J. Tang, L. Jian, and Z. Yang, "FlipDA: Effective and robust data augmentation for few-shot learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 8646–8665.

[16] S. Ghosh, C. K. R. Evuru, S. Kumar, R. S, S. Sakshi, U. Tyagi, and D. Manocha, "DALE: generative data augmentation for low-resource legal NLP," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 8511–8565.

[17] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 271:1–271:40, 2023.

[18] J. Wang, C. Wang, J. Huang, M. Gao, and A. Zhou, "Uncertainty-aware self-training for low-resource neural sequence labeling," in *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 2023, pp. 13 682–13 690.

[19] V. Kumar, A. Choudhary, and E. Cho, "Data augmentation using pre-trained transformer models," *CoRR*, vol. abs/2003.02245, 2020.

[20] C. Xia, C. Zhang, H. Nguyen, J. Zhang, and P. S. Yu, "CG-BERT: Conditional text generation with BERT for generalized few-shot intent detection," *CoRR*, vol. abs/2004.01881, 2020.

[21] C. Xia, C. Xiong, and P. S. Yu, "Pseudo siamese network for few-shot intent generation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2005–2009.

[22] N. Malandrakis, M. Shen, A. K. Goyal, S. Gao, A. Sethi, and A. Metallinou, "Controlled text generation for data augmentation in intelligent artificial agents," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 2019, pp. 90–98.

[23] T. Dopierre, C. Gravier, and W. Logerais, "ProtAugment: Intent detection meta-learning through unsupervised diverse paraphrasing," in *roceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 2454–2466.

[24] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, and L. Kong, "ZeroGen: Efficient zero-shot learning via dataset generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11 653–11 669.

[25] J. Ye, N. Xu, Y. Wang, J. Zhou, Q. Zhang, T. Gui, and X. Huang, "LLM-DA: Data augmentation via large language models for few-shot named entity recognition," *CoRR*, vol. abs/2402.14568, 2024.

[26] N. Popovic, A. Y. Kangen, T. Schopf, and M. Färber, "DocIE@XLLM25: In-context learning for information extraction using fully synthetic demonstrations," in *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling*, 2025, pp. 298–309.

[27] L. Peng, Y. Zhang, and J. Shang, "Controllable data augmentation for few-shot text mining with chain-of-thought attribute manipulation," in *Findings of the Association for Computational Linguistics: ACL*, 2024, pp. 1–16.

[28] W. Wu, W. Li, J. Liu, X. Xiao, S. Li, and Y. Lyu, "Precisely the point: Adversarial augmentations for faithful and informative text generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,*, 2022, pp. 7160–7176.

[29] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[30] A. M. Saxe, Y. Bansal, J. Dapello, M. S. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, 2018.

[31] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[32] S. Chen, E. Dobriban, and J. H. Lee, "A group-theoretic framework for data augmentation," in *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, 2020.

[33] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[34] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, 2017, pp. 4080–4090.

[35] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, 2016, pp. 3637–3645.

[36] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H. Hon, "Unified language model pre-training for natural language understanding and generation," in *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*, 2019, pp. 13 042–13 054.

[37] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, 2020, pp. 7871–7880.

[38] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, "An evaluation dataset for intent classification and out-of-scope prediction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 1311–1316.

[39] W. He, Y. Dai, Y. Zheng, Y. Wu, Z. Cao, D. Liu, P. Jiang, M. Yang, F. Huang, L. Si, J. Sun, and Y. Li, "GALAXY: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022, pp. 10 749–10 757.

[40] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 110–119.

[41] J. Su, "Bert4keras," https://bert4keras.spaces.ac.cn, 2020.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the 7th International Conference on Learning Representations*, 2019.