Joint Sequence–Vocabulary Selection for Efficient LLM Distillation

Xueli Geng[1] **Weicheng Zhao**[2] **Xutong Mu**[3,†]
Tianrui Wei[3] Yanbiao Ma[4] Yulong Shen[3]
[1]School of Artificial Intelligence, Xidian University
[2]School of Economics and Management, Xidian University
[3]School of Computer Science and Technology, Xidian University
[4]Gaoling School of Artificial Intelligence, Renmin University of China
muxutong@xidian.edu.cn
[†]Corresponding author.

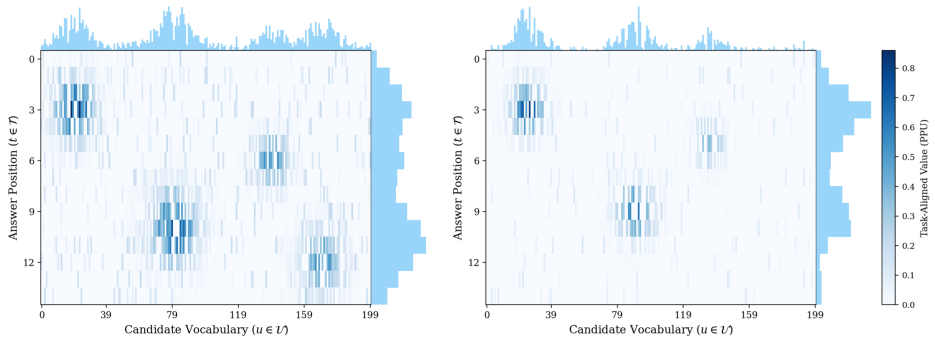
Abstract

Selective knowledge distillation has become increasingly important for compressing large language models, where dense logit-based knowledge transfer over long sequences and large vocabularies incurs substantial computational overhead. Distillation supervision is naturally organized over a sequence–vocabulary matrix, with its highest-information components often concentrated on a small subset of specific sequence–vocabulary pairs. Existing methods typically select sequence positions or vocabulary candidates independently, relying on coarse-grained one-dimensional heuristics that weaken the precision and effectiveness of knowledge transfer. To address this, we propose **J**oint **S**equence–**V**ocabulary **D**istillation (**JSVD**), a plug-and-play selective distillation framework that identifies and focuses supervision on informative sequence–vocabulary pairs. For each training example, JSVD constructs bidirectional teacher–student candidate supports, scores sequence–vocabulary pairs using prediction discrepancies, and progressively focuses distillation on unresolved informative pairs during training. Extensive experiments across diverse distillation objectives, model pairs, and benchmarks show that JSVD consistently improves downstream performance while reducing distillation overhead and accelerating convergence.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks [1–5]. However, their substantial computational, memory, and energy costs pose significant challenges for deployment in resource-constrained and latency-sensitive scenarios. Knowledge Distillation (KD) [6] provides a practical way to compress LLMs by transferring knowledge from a powerful teacher model to a smaller student model. In particular, logit-based distillation transfers richer supervision by matching the teacher's output distribution rather than relying solely on hard labels [7, 8]. Nevertheless, for autoregressive LLMs, such distribution matching must be performed across both long token sequences and large vocabularies, leading to substantial computational and memory overhead. Consequently, selective distillation has become an important strategy for improving the efficiency of LLM compression by focusing training on the most informative teacher signals [9, 10].

Existing selective distillation methods mainly sparsify logit-based supervision along either the sequence or vocabulary axis. Sequence-side methods [11–13] select important token positions, typically according to student difficulty or teacher confidence, but still retain full-vocabulary supervision at selected positions. Vocabulary-side methods [14, 9] reduce distribution-matching cost by keep-



(a) Early training phase (a) Post-training phase

Figure 1: Task-aligned distillation value over the sequence–vocabulary plane at early and late training stages. Darker cells indicate higher value for distilling candidate u at position t, with marginal histograms showing aggregated values along each axis.

ing informative vocabulary entries, such as top-k logits, yet usually apply position-independent selection rules. As a result, both paradigms treat selection in a dimension-wise manner and fail to capture a key property of autoregressive distillation: the usefulness of vocabulary-level supervision is position-dependent, requiring different informative candidates at different sequence positions.

A more fine-grained perspective is to view logit-based distillation for autoregressive LLMs as supervision over a sequence–vocabulary matrix, where each entry reflects the supervision signal associated with vocabulary candidate u at sequence position t. As shown in Figure 1, high-value distillation signals are sparse and localized, concentrating on specific sequence–vocabulary pairs rather than entire sequence positions or globally important vocabulary candidates. This suggests that sequence-side selection may retain redundant vocabulary supervision, while vocabulary-side selection based on distributional heuristics may miss position-dependent pair-level information. Moreover, as the student evolves during training, the utility of these pairs changes, requiring distillation to adaptively refocus on unresolved informative signals. These observations motivate a joint and dynamic formulation of selective distillation that operates directly at the sequence–vocabulary pair level.

Motivated by this insight, we propose **Joint Sequence–Vocabulary Distillation** (**JSVD**), a plug-and-play selective distillation framework for LLMs. Instead of making separate sequence-side or vocabulary-side decisions, JSVD selects supervision directly in the sequence–vocabulary space. For each training example, it builds bidirectional teacher–student candidate supports, scores candidate pairs by prediction discrepancy, and progressively focuses on unresolved high-value pairs as the student evolves. JSVD is compatible with multiple logit-based objectives, including FKL, RKL, AKL, and ToDi. Extensive experiments show that JSVD consistently improves downstream performance, reduces distillation overhead, and accelerates convergence over full and dimension-decoupled selective distillation baselines.

The main contributions are summarized as follows:

- We formulate selective logit distillation as a joint sequence–vocabulary pair selection problem, revealing a fine-grained supervision structure beyond sequence-only or vocabulary-only selection.
- We propose JSVD, a plug-and-play framework that integrates bidirectional teacher–student candidate construction, discrepancy-based pair scoring, and adaptive focusing to allocate distillation supervision more precisely and efficiently.
- We provide comprehensive empirical validation across KD objectives, teacher–student model pairs, and benchmarks, showing that JSVD improves downstream performance while reducing overhead and accelerating convergence.

2 Related Work

Knowledge Distillation for LLMs. Knowledge Distillation[6] (KD) is a technique where a large teacher model guides a smaller student model to learn from its knowledge [15, 16, 8]. This can be classified into white-box distillation [7, 17] and black-box distillation [18–20]. White-box distillation, such as logits-based matching, aligns the teacher and student models by comparing their logits

directly [6]. The core objective functions in KD are forward KL divergence(FKL) [6] and reverse KL divergence(RKL) [7]. Building on these, several improved variants have been proposed, including AKL [21] and ToDi [22], which dynamically adjust FKL and RKL to optimize the distillation process. ABKD [23] introduces flexibility by controlling the mode preference with hyperparameters.

Selective Token Distillation. Selective token distillation aims to reduce the cost of logit-based KD by applying distillation only to informative parts of the teacher distribution. Existing methods mainly sparsify the distillation signal along two axes: sequence positions and vocabulary candidates. Sequence-position selection focuses the distillation loss on a subset of important positions, typically according to student learning difficulty, teacher confidence, or task-specific token filtering [13, 11, 24]. While this reduces redundant supervision across the sequence, selected positions are often still distilled with full-vocabulary distributions, leaving substantial vocabulary-level redundancy. Vocabulary-side selection instead restricts the candidate support at each position, for example through top-k logits, ranking-based objectives, or sampling-based approximations [14, 9, 25, 10]. These methods improve memory and computational efficiency, but usually apply a fixed or locally defined vocabulary rule at each position, without modeling how the required vocabulary support varies across sequence positions. Recent multi-axis selective KD studies [26] further combine selection along position, class, and sample axes, but their decisions remain axis-level. In contrast, JSVD selects individual sequence–vocabulary pairs (t, u), directly capturing which vocabulary candidates remain informative at each specific sequence position.

3 Preliminaries

Autoregressive LLMs generate tokens sequentially. Given an input sequence $\mathbf{x}$, at each step t, the teacher and student define conditional distributions over the vocabulary $\mathcal{V}$, where $p_t(u) = p(u \mid \mathbf{x}, \mathbf{y}_{<t})$ and $q_t(u) = q_\theta(u \mid \mathbf{x}, \mathbf{y}_{<t})$ for $u \in \mathcal{V}$. For each sequence–vocabulary pair (t, u), the forward-KL contribution and the full logit distillation loss are written as

$$D_{\mathrm{FKL}}^{(t,u)}(p, q_\theta) = p_t(u) \log \frac{p_t(u)}{q_t(u)}, \qquad \mathcal{L}_{\mathrm{FKL}} = \sum_{(t,u)\in[L]\times\mathcal{V}} D_{\mathrm{FKL}}^{(t,u)}(p, q_\theta).$$

Sequence- and vocabulary-side selective distillation. Sequence-side methods select a subset of positions $S \subseteq [L]$ and retain full-vocabulary supervision at these positions, while vocabulary-side methods select a candidate subset $U_t \subseteq \mathcal{V}$, e.g., $U_t = \mathrm{TopK}(p_t)$, at each position. Their losses can be written as

$$\mathcal{L}_{\mathrm{seqKL}} = \sum_{t\in S} \sum_{u\in\mathcal{V}} D_{\mathrm{FKL}}^{(t,u)}(p, q_\theta), \qquad \mathcal{L}_{\mathrm{vocabKL}} = \sum_{t=1}^{L} \sum_{u\in U_t} D_{\mathrm{FKL}}^{(t,u)}(p, q_\theta).$$

4 Methodology

Given the pair-level selected support $\Omega_s(x) \subseteq [L] \times \mathcal{V}$, we define the induced vocabulary support at position t as $\mathcal{A}_{s,t}(x) = \{u \in \mathcal{V}_t^s \mid (t, u) \in \Omega_s(x)\}$. The active sequence positions are then determined by $\mathcal{S}_s(x) = \{t \in [L] \mid \mathcal{A}_{s,t}(x) \neq \emptyset\}$. Importantly, both supports are induced by the same pair-level selection, rather than by independent sequence- and vocabulary-side decisions. Positions with $\mathcal{A}_{s,t}(x) = \emptyset$ are excluded from distillation.

Conceptually, JSVD restricts distillation supervision to the selected sequence–vocabulary pairs:

$$\mathcal{L}_{\mathrm{JSVD}}^{\mathrm{concept}} = \sum_{(t,u)\in\Omega_s(x)} D^{(t,u)}(p_t, q_{\theta,t}). \tag{1}$$

Here, $D^{(t,u)}$ denotes a conceptual pair-level divergence contribution, associated with objectives such as FKL [6], RKL [7], AKL [21], and ToDi [22]. This formulation specifies where supervision is applied, while the actual training objective, defined in Eq. (10), incorporates support-dependent renormalization. Thus, $\Omega_s(x)$ determines the selected pairs, whereas the renormalized distributions determine their distillation contributions.

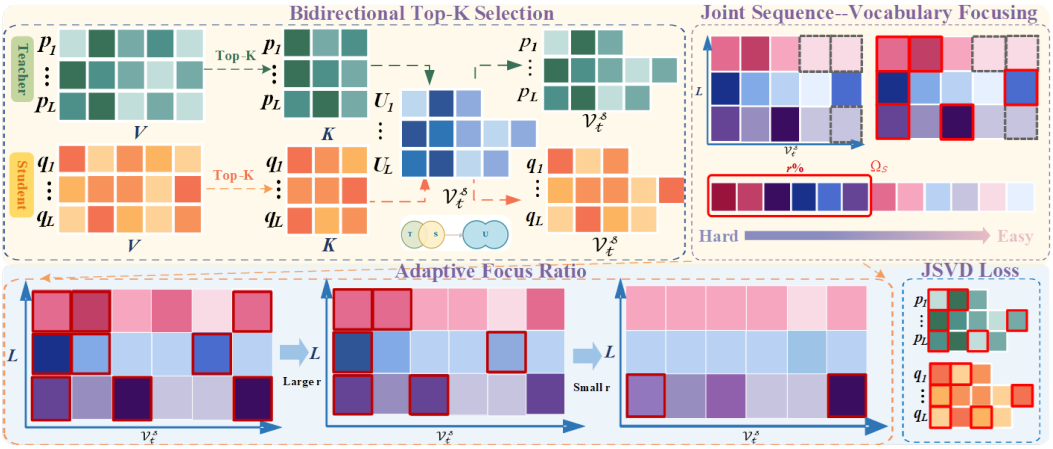


Figure 2: Overview of JSVD, which selects informative sequence–vocabulary pairs for adaptive logit distillation.

4.1 Bidirectional Top-K Selection

Before performing pair-level selection, JSVD first constructs a compact candidate support at each sequence position. This step is motivated by the long-tailed nature of LLM output distributions: most vocabulary entries have near-zero probabilities and contribute little useful supervision, while retaining the full vocabulary would make pair-level scoring prohibitively expensive [9, 10]. We therefore use top-k selection as a high-recall candidate construction step rather than as the final selection rule.

Relying only on the teacher top-k candidates may miss high-confidence student errors, while relying only on the student top-k candidates may discard informative teacher signals. Thus, at each sequence position t, we select the top-k vocabulary candidates from both the teacher and student distributions:

$$\mathcal{I}_t^p = \mathrm{Top}_k(p_t), \qquad \mathcal{I}_t^q = \mathrm{Top}_k(q_t), \tag{2}$$

where $\mathcal{I}_t^p$ and $\mathcal{I}_t^q$ denote the teacher-side and student-side top-k candidate index sets, respectively. The bidirectional candidate support is then defined as

$$\mathcal{V}_t^s = \mathcal{I}_t^p \cup \mathcal{I}_t^q. \tag{3}$$

with $k \le |\mathcal{V}_t^s| \le 2k$. We then gather the teacher and student probabilities on this support:

$$\tilde{p}_t = p_t[\mathcal{V}_t^s], \quad \tilde{q}_t = q_t[\mathcal{V}_t^s]. \tag{4}$$

This step only defines the candidate sequence–vocabulary pairs $\{(t,u) : u \in \mathcal{V}_t^s\}$; the final informative pairs are selected by the subsequent joint sequence-vocabulary focusing operation.

4.2 Joint Sequence–Vocabulary Focusing

After constructing the bidirectional candidate support, JSVD performs fine-grained selection over sequence–vocabulary pairs. For each training example, we define the candidate pair set as

$$\mathcal{C}(\mathbf{x}) = \{(t,u) \mid t \in \{1,\dots,L\},\ u \in \mathcal{V}_t^s\}, \tag{5}$$

where $\mathcal{V}_t^s$ is the bidirectional candidate support at sequence position t. For each candidate pair $(t,u) \in \mathcal{C}(\mathbf{x})$, we compute a local difficulty score using the teacher–student prediction discrepancy:

$$H(t,u) = |p_t(u) - q_t(u)|\,, \tag{6}$$

where $p_t(u) = p(u \mid \mathbf{x}, \mathbf{y}_{<t})$ and $q_t(u) = q_\theta(u \mid \mathbf{x}, \mathbf{y}_{<t})$. This score measures the unresolved mismatch between teacher and student at a specific sequence–vocabulary pair.

Given a focus ratio r_s at training step s, JSVD selects the top fraction of candidate pairs with the largest difficulty scores:

$$\Omega_s(\mathbf{x}) = \mathrm{Top}_{r_s}\left(\{H(t,u) \mid (t,u) \in \mathcal{C}(\mathbf{x})\}\right), \tag{7}$$

where $\Omega_s(\mathbf{x}) \subseteq \mathcal{C}(\mathbf{x})$ denotes the selected sequence–vocabulary pairs. Equivalently, the selected vocabulary support at position t is

$$\mathcal{A}_t = \{u \mid (t, u) \in \Omega_s(\mathbf{x})\}. \tag{8}$$

The size of $\mathcal{A}_t$ is not fixed across positions and can be zero for low-value positions. Thus, JSVD allocates supervision according to global pair-level difficulty rather than assigning the same vocabulary budget to every sequence position.

For each position with $\mathcal{A}_t \neq \emptyset$, we renormalize the teacher and student probabilities over the selected support:

$$\bar{p}_t(u) = \frac{p_t(u)}{\sum_{v \in \mathcal{A}_t} p_t(v)}, \qquad \bar{q}_t(u) = \frac{q_t(u)}{\sum_{v \in \mathcal{A}_t} q_t(v)}, \quad u \in \mathcal{A}_t. \tag{9}$$

The resulting pair-level selective distillation loss is then computed only over the selected pairs:

$$\mathcal{L}_{\mathrm{JSVD}} = \sum_{t:\mathcal{A}_t \neq \emptyset} \sum_{u \in \mathcal{A}_t} D^{(t,u)}(\bar{p}_t, \bar{q}_t). \tag{10}$$

4.3 Adaptive Focus Ratio

Although JSVD identifies informative sequence–vocabulary pairs, the number of selected pairs should adapt to the evolving student state. A fixed focus ratio may either retain redundant supervision or discard useful signals. We therefore adjust the focus ratio using the teacher–student discrepancy over candidate pairs.

For a mini-batch $\mathcal{B}_s$ at step s, the average discrepancy difficulty is defined as

$$\bar{D}_s = \frac{1}{|\mathcal{B}_s|} \sum_{\mathbf{x} \in \mathcal{B}_s} \frac{1}{|\mathcal{C}_s(\mathbf{x})|} \sum_{(t,u) \in \mathcal{C}_s(\mathbf{x})} \left| p_t(u) - q_{\theta_s,t}(u) \right|, \tag{11}$$

where $\mathcal{C}_s(\mathbf{x})$ is the candidate pair set constructed in Section 4.1.

We compare $\bar{D}_s$ with the previous EMA reference D_{s-1}^{ref} and compute the relative discrepancy change:

$$e_s = \frac{\bar{D}_s - D_{s-1}^{\mathrm{ref}}}{D_{s-1}^{\mathrm{ref}} + \epsilon}. \tag{12}$$

The focus ratio is then updated by

$$r_s = \mathrm{clip}_{[0,1]} \left(r_{s-1} + \alpha e_s \right), \tag{13}$$

where α controls the adaptation speed and ϵ ensures numerical stability. Finally, the EMA reference is updated as

$$D_s^{\mathrm{ref}} = \beta D_{s-1}^{\mathrm{ref}} + (1 - \beta) \bar{D}_s, \tag{14}$$

where β controls the smoothing strength.

When $\bar{D}_s$ exceeds the reference, r_s increases to expand supervision; otherwise, it decreases to focus on harder pairs. Given r_s, JSVD selects the top-r_s fraction of candidate pairs by $H_s(t, u)$ to form $\Omega_s(\mathbf{x})$, progressively concentrating supervision on unresolved high-discrepancy pairs during training.

4.4 Theoretical Justification for Discrepancy-Based Pair Scoring

JSVD scores each sequence–vocabulary pair (t, u) by the teacher–student discrepancy

$$H_s(t, u) = |p_t(u) - q_{\theta_s,t}(u)|.$$

This scoring rule is applied before support-dependent renormalization, whereas the selective KL loss is evaluated on the renormalized distributions over the selected support A_t. To connect these two quantities, Appendix B shows that, when A_t has non-negligible probability mass and small teacher–student mass mismatch, the raw discrepancy $|p_t(u) - q_t(u)|$ and the restricted-support discrepancy $|p_t^A(u) - q_t^A(u)|$ are locally order-equivalent. Therefore, the raw discrepancy used by JSVD can be viewed as a stable proxy for the restricted-support optimization signal analyzed below.

Proposition 1 (Discrepancy is an objective-agnostic local distillation signal)**.** *Let p_t and q_t be normalized teacher and student distributions on a fixed support, and define $\Delta_t = q_t - p_t$. For any twice continuously differentiable proper distillation objective $D(p_t, q_t)$ whose minimum is attained at $q_t = p_t$, there exists a matrix G_{p_t} that is positive definite on the simplex tangent space such that, as $q_t \to p_t$,*

$$D(p_t, q_t) = \frac{1}{2}\Delta_t^\top G_{p_t}\Delta_t + o(\|\Delta_t\|_2^2),$$

and its logit gradient satisfies

$$\nabla_{z_t} D(p_t, q_t) = J(p_t)G_{p_t}\Delta_t + o(\|\Delta_t\|_2),$$

where $J(p_t)$ is the softmax Jacobian. Consequently, the teacher–student discrepancy Δ_t determines the leading-order local loss and optimization direction for smooth KL-based distillation objectives.

Second, pair-level selection captures structures that axis-wise selection cannot recover. Aggregating discrepancies along the sequence or vocabulary axis loses the positional dependence of vocabulary-level errors. As a result, two examples can share identical sequence-wise and vocabulary-wise aggregate scores while requiring different optimal pair supports. This motivates selecting directly in the sequence–vocabulary space rather than composing independent sequence and vocabulary selections. Proposition 1 shows that large-discrepancy pairs correspond to strong local distillation signals. Under FKL, the discrepancy exactly equals the coordinate-wise logit-gradient magnitude; under RKL, AKL, and ToDi, it remains a leading-order proxy for the same optimization signal.

Proposition 2 (Axis-wise selection cannot recover pair-level structure)**.** *There exist sequence–vocabulary discrepancy matrices with identical sequence-wise and vocabulary-wise aggregate scores but different optimal top-r pair supports. Therefore, selectors based only on sequence or vocabulary aggregates can miss localized high-value distillation pairs.*

Proposition 2 formalizes the limitation of dimension-decoupled selection. A vocabulary candidate may be informative at one position but redundant at another, so high-value supervision is generally tied to specific (t, u) pairs. JSVD preserves this coupling by ranking and selecting pairs directly in the sequence–vocabulary space. The full proofs are provided in Appendix B.

5 Experiment

5.1 Experimental Setup

Training Configuration. All experiments are conducted with open-source local models on eight NVIDIA H100 80GB GPUs. All methods, including the supervised fine-tuning baseline and distillation variants, optimize the student model with the same LoRA configuration [27]: rank 16, scaling factor 32, and dropout 0. We evaluate three teacher–student pairs: Qwen2.5-7B→Qwen2.5-3B, OpenLLaMA2-7B→OpenLLaMA2-3B, and Qwen2.5-7B→Qwen2.5-1.5B. And all experiments are conducted with five different random seeds, and results are reported as mean ± standard deviation. The best scores are highlighted in bold, and the second-best scores are underlined. The full experimental setup and default hyperparameter settings are summarized in Table 7, and the complete JSVD training procedure is provided in Algorithm 1 in Appendix C and D.

Tasks and Evaluation. For general instruction-following, we train on the OLMo2 version of Tulu [28, 29] and evaluate after supervised fine-tuning and distillation on DollyEval [30], SelfInst [31], VicunaEval [32], S-NI [33], and UnNI [34], using ROUGE-L as the metric. For mathematical reasoning, the student is first fine-tuned on the full MetaMathQA dataset for one epoch, then distilled on 50k MetaMathQA samples [35], and evaluated on GSM8K [36] and MATH [37]. For code generation, the student is trained for two epochs with WizardCoder Evol-Instruct prompts [38] and evaluated on HumanEval [39] and MBPP [40]. We report exact-match accuracy for mathematical reasoning and Pass@1 for code generation.

Baseline Methods. We adopt four representative logit-based KD objectives as backbones: FKL [6], RKL [7], AKL [21], and ToDi [22]. To evaluate selection-based distillation, we consider three representative strategies. (i) Vocabulary selection: Top-K [9] (Top-200), which retains the highest-probability teacher tokens per step. (ii) Sequence selection: AdaKD [41], which selects or reweights

Table 1: Comparison of ROUGE-L scores on five instruction-following benchmarks under the Qwen2.5-7B→Qwen2.5-3B setting.

Method	DollyEval	S-NI	UnNI	SelfInst	VicunaEval	Avg.
			Qwen2.5-7B → Qwen2.5-3B			
Teacher(Lora)	$30.69_{\pm 0.14}$	$45.65_{\pm 0.11}$	$55.37_{\pm 0.68}$	$28.67_{\pm 0.03}$	$20.99_{\pm 0.86}$	36.27
Student(Lora)	$24.96_{\pm 0.50}$	$38.79_{\pm 0.38}$	$40.62_{\pm 0.54}$	$22.13_{\pm 0.16}$	$15.63_{\pm 0.54}$	28.43
FKL	$24.98_{\pm 0.41}$	$39.16_{\pm 0.16}$	$42.27_{\pm 0.22}$	$24.56_{\pm 0.30}$	$16.53_{\pm 0.74}$	29.50
FKL+Top-200	$24.45_{\pm 0.36}$	$39.00_{\pm 0.88}$	$42.28_{\pm 0.23}$	$24.30_{\pm 0.17}$	$16.00_{\pm 0.74}$	29.21
FKL+AdaKD	$25.33_{\pm 0.34}$	$39.40_{\pm 0.57}$	$42.55_{\pm 0.51}$	$24.77_{\pm 0.30}$	$17.01_{\pm 0.74}$	29.81
FKL+indep-2D	$25.20_{\pm 0.30}$	$39.21_{\pm 0.50}$	$42.60_{\pm 0.10}$	$24.80_{\pm 0.32}$	$17.00_{\pm 0.70}$	29.76
FKL+JSVD	$26.65_{\pm 0.81}$	$40.45_{\pm 0.77}$	$43.69_{\pm 0.84}$	$27.46_{\pm 0.28}$	$18.18_{\pm 0.35}$	31.29
RKL	$27.84_{\pm 0.60}$	$40.56_{\pm 0.59}$	$44.89_{\pm 0.18}$	$26.49_{\pm 0.70}$	$20.06_{\pm 0.33}$	31.97
RKL+Top-200	$27.71_{\pm 0.50}$	$40.54_{\pm 0.66}$	$44.64_{\pm 0.18}$	$26.39_{\pm 0.70}$	$19.67_{\pm 0.31}$	31.79
RKL+AdaKD	$28.36_{\pm 0.40}$	$41.78_{\pm 0.80}$	$46.11_{\pm 0.18}$	$26.57_{\pm 0.70}$	$20.61_{\pm 0.53}$	32.68
RKL+indep-2D	$27.93_{\pm 0.40}$	$41.78_{\pm 0.60}$	$46.23_{\pm 0.10}$	$26.13_{\pm 0.17}$	$20.34_{\pm 0.30}$	32.48
RKL+JSVD	$29.96_{\pm 0.54}$	$42.43_{\pm 0.30}$	$47.57_{\pm 0.17}$	$\mathbf{27.77}_{\pm 0.87}$	$20.53_{\pm 0.68}$	33.65
AKL	$29.55_{\pm 0.28}$	$41.11_{\pm 0.81}$	$45.13_{\pm 0.67}$	$21.24_{\pm 0.03}$	$15.81_{\pm 0.63}$	30.57
AKL+Top-200	$29.37_{\pm 0.44}$	$41.10_{\pm 0.70}$	$45.01_{\pm 0.42}$	$21.18_{\pm 0.50}$	$15.70_{\pm 0.34}$	30.47
AKL+AdaKD	$30.72_{\pm 0.40}$	$42.04_{\pm 0.80}$	$46.65_{\pm 0.60}$	$21.55_{\pm 0.50}$	$16.44_{\pm 0.40}$	31.48
AKL+indep-2D	$30.80_{\pm 0.30}$	$42.00_{\pm 0.80}$	$46.90_{\pm 0.60}$	$21.53_{\pm 0.50}$	$16.70_{\pm 0.40}$	31.58
AKL+JSVD	$\mathbf{32.77}_{\pm 0.65}$	$\mathbf{43.16}_{\pm 0.79}$	$47.74_{\pm 0.64}$	$22.66_{\pm 0.56}$	$16.79_{\pm 0.24}$	32.62
ToDi	$28.79_{\pm 0.66}$	$42.63_{\pm 0.19}$	$48.79_{\pm 0.74}$	$25.78_{\pm 0.66}$	$21.73_{\pm 0.93}$	33.54
ToDi+Top-200	$28.81_{\pm 0.60}$	$42.50_{\pm 0.10}$	$48.67_{\pm 0.60}$	$25.63_{\pm 0.57}$	$21.60_{\pm 0.81}$	33.44
ToDi+AdaKD	$29.60_{\pm 0.66}$	$42.54_{\pm 0.30}$	$49.30_{\pm 0.40}$	$26.25_{\pm 0.60}$	$22.59_{\pm 0.40}$	34.06
ToDi+indep-2D	$29.20_{\pm 0.60}$	$42.30_{\pm 0.20}$	$48.97_{\pm 0.63}$	$26.34_{\pm 0.10}$	$22.40_{\pm 0.61}$	33.85
ToDi+JSVD	$31.19_{\pm 0.79}$	$43.03_{\pm 0.84}$	$\mathbf{49.66}_{\pm 0.51}$	$26.84_{\pm 0.32}$	$\mathbf{23.14}_{\pm 0.73}$	**34.77**

Table 2: Comparison of ROUGE-L scores on five instruction-following benchmarks across additional teacher–student settings.

	OpenLLaMA2-7B → OpenLLaMA2-3B						**Qwen2.5-7B → Qwen2.5-1.5B**					
Method	**Dolly**	**S-NI**	**UnNI**	**SelfInst**	**Vicuna**	**Avg.**	**Dolly**	**S-NI**	**UnNI**	**SelfInst**	**Vicuna**	**Avg.**
Teacher(Lora)	$27.36_{\pm 0.34}$	$36.96_{\pm 0.11}$	$33.13_{\pm 0.35}$	$20.41_{\pm 0.88}$	$18.96_{\pm 0.77}$	27.36	$30.69_{\pm 0.14}$	$45.65_{\pm 0.11}$	$55.37_{\pm 0.68}$	$28.67_{\pm 0.03}$	$20.99_{\pm 0.86}$	36.27
Student(Lora)	$24.10_{\pm 0.84}$	$32.79_{\pm 0.69}$	$30.69_{\pm 0.33}$	$17.63_{\pm 0.56}$	$16.01_{\pm 0.13}$	24.24	$17.82_{\pm 0.30}$	$30.10_{\pm 0.21}$	$33.54_{\pm 0.20}$	$18.30_{\pm 0.11}$	$13.02_{\pm 0.30}$	22.56
FKL	$22.93_{\pm 0.69}$	$33.79_{\pm 0.44}$	$31.54_{\pm 0.79}$	$17.77_{\pm 0.99}$	$16.96_{\pm 0.41}$	24.60	$17.31_{\pm 0.80}$	$31.01_{\pm 0.14}$	$34.30_{\pm 0.90}$	$19.98_{\pm 0.30}$	$15.53_{\pm 0.20}$	23.63
FKL+Top-200	$22.95_{\pm 0.40}$	$33.70_{\pm 0.40}$	$31.50_{\pm 0.61}$	$17.76_{\pm 0.39}$	$16.93_{\pm 0.34}$	24.57	$16.90_{\pm 0.19}$	$30.51_{\pm 0.27}$	$34.00_{\pm 0.60}$	$19.38_{\pm 0.17}$	$15.01_{\pm 0.10}$	23.16
FKL+AdaKD	$23.10_{\pm 0.53}$	$33.90_{\pm 0.76}$	$32.30_{\pm 0.30}$	$18.51_{\pm 0.45}$	$17.53_{\pm 0.50}$	25.07	$18.53_{\pm 0.40}$	$32.30_{\pm 0.90}$	$35.14_{\pm 0.40}$	$21.03_{\pm 0.34}$	$17.20_{\pm 0.40}$	24.84
FKL+Indep-2D	$23.14_{\pm 0.50}$	$33.46_{\pm 0.60}$	$32.10_{\pm 0.30}$	$18.20_{\pm 0.50}$	$17.67_{\pm 0.47}$	24.91	$18.23_{\pm 0.20}$	$32.43_{\pm 0.70}$	$35.02_{\pm 0.31}$	$20.83_{\pm 0.40}$	$17.57_{\pm 0.20}$	24.82
FKL+JSVD	$23.56_{\pm 0.71}$	$34.04_{\pm 0.71}$	$33.16_{\pm 0.83}$	$18.93_{\pm 0.84}$	$18.03_{\pm 0.53}$	25.54	$18.90_{\pm 0.20}$	$33.10_{\pm 0.30}$	$35.67_{\pm 0.38}$	$21.23_{\pm 0.10}$	$17.43_{\pm 0.30}$	25.27
RKL	$24.12_{\pm 0.37}$	$34.67_{\pm 0.92}$	$33.59_{\pm 0.11}$	$20.00_{\pm 0.69}$	$18.69_{\pm 0.30}$	26.21	$20.14_{\pm 0.40}$	$31.69_{\pm 0.20}$	$35.91_{\pm 0.20}$	$21.49_{\pm 0.10}$	$16.61_{\pm 0.40}$	25.17
RKL+Top-200	$24.01_{\pm 0.10}$	$34.60_{\pm 0.92}$	$33.50_{\pm 0.10}$	$19.80_{\pm 0.40}$	$18.63_{\pm 0.30}$	26.11	$19.93_{\pm 0.40}$	$31.20_{\pm 0.20}$	$35.00_{\pm 0.30}$	$21.60_{\pm 0.40}$	$16.79_{\pm 0.40}$	24.90
RKL+AdaKD	$24.57_{\pm 0.20}$	$35.44_{\pm 0.40}$	$\mathbf{34.75}_{\pm 0.20}$	$20.91_{\pm 0.40}$	$19.59_{\pm 0.31}$	27.05	$21.94_{\pm 0.40}$	$32.78_{\pm 0.80}$	$37.21_{\pm 0.34}$	$23.57_{\pm 0.40}$	$18.80_{\pm 0.30}$	26.86
RKL+Indep-2D	$24.11_{\pm 0.20}$	$35.50_{\pm 0.40}$	$33.93_{\pm 0.20}$	$20.51_{\pm 0.40}$	$19.43_{\pm 0.31}$	26.70	$21.34_{\pm 0.40}$	$32.80_{\pm 0.80}$	$37.00_{\pm 0.34}$	$23.34_{\pm 0.40}$	$18.33_{\pm 0.30}$	26.56
RKL+JSVD	$25.19_{\pm 0.33}$	$36.84_{\pm 0.63}$	$34.64_{\pm 0.46}$	$\mathbf{21.21}_{\pm 0.69}$	$19.70_{\pm 0.85}$	27.52	$\mathbf{23.10}_{\pm 0.70}$	$33.90_{\pm 0.20}$	$\mathbf{38.90}_{\pm 0.30}$	$\mathbf{25.34}_{\pm 0.60}$	$19.42_{\pm 0.30}$	**28.13**
AKL	$25.43_{\pm 0.66}$	$34.64_{\pm 0.15}$	$32.49_{\pm 0.08}$	$18.54_{\pm 0.79}$	$18.79_{\pm 0.53}$	25.98	$19.30_{\pm 0.20}$	$31.87_{\pm 0.30}$	$35.50_{\pm 0.50}$	$17.60_{\pm 0.14}$	$16.30_{\pm 0.40}$	24.11
AKL+Top-200	$25.40_{\pm 0.30}$	$34.59_{\pm 0.20}$	$32.50_{\pm 0.36}$	$18.54_{\pm 0.70}$	$18.93_{\pm 0.40}$	25.99	$19.27_{\pm 0.10}$	$30.69_{\pm 0.30}$	$35.63_{\pm 0.40}$	$20.41_{\pm 0.20}$	$16.00_{\pm 0.40}$	24.40
AKL+AdaKD	$26.10_{\pm 0.10}$	$35.37_{\pm 0.30}$	$33.27_{\pm 0.41}$	$19.30_{\pm 0.60}$	$19.70_{\pm 0.50}$	26.75	$20.10_{\pm 0.20}$	$32.60_{\pm 0.22}$	$36.30_{\pm 0.61}$	$21.38_{\pm 0.20}$	$19.26_{\pm 0.57}$	25.93
AKL+Indep-2D	$25.70_{\pm 0.10}$	$35.00_{\pm 0.20}$	$33.43_{\pm 0.80}$	$19.01_{\pm 0.76}$	$19.20_{\pm 0.50}$	26.47	$19.79_{\pm 0.10}$	$32.41_{\pm 0.40}$	$36.04_{\pm 0.13}$	$21.50_{\pm 0.34}$	$19.03_{\pm 0.70}$	25.75
AKL+JSVD	$26.56_{\pm 0.09}$	$35.72_{\pm 0.19}$	$33.50_{\pm 0.21}$	$19.47_{\pm 0.30}$	$19.66_{\pm 0.41}$	26.98	$20.68_{\pm 0.20}$	$33.44_{\pm 0.10}$	$36.90_{\pm 0.40}$	$22.16_{\pm 0.56}$	$19.29_{\pm 0.24}$	26.49
ToDi	$26.63_{\pm 0.02}$	$35.97_{\pm 0.61}$	$32.86_{\pm 0.27}$	$19.67_{\pm 0.11}$	$18.79_{\pm 0.07}$	26.78	$19.43_{\pm 0.60}$	$32.05_{\pm 0.10}$	$34.60_{\pm 0.13}$	$19.60_{\pm 0.30}$	$16.53_{\pm 0.10}$	24.44
ToDi+Top-200	$26.57_{\pm 0.10}$	$35.90_{\pm 0.71}$	$32.79_{\pm 0.20}$	$19.67_{\pm 0.13}$	$18.70_{\pm 0.10}$	26.73	$19.00_{\pm 0.40}$	$31.45_{\pm 0.10}$	$34.11_{\pm 0.50}$	$19.50_{\pm 0.50}$	$16.54_{\pm 0.10}$	24.12
ToDi+AdaKD	$27.11_{\pm 0.50}$	$36.50_{\pm 0.30}$	$33.40_{\pm 0.10}$	$20.20_{\pm 0.22}$	$19.40_{\pm 0.20}$	27.32	$20.60_{\pm 0.20}$	$32.93_{\pm 0.30}$	$36.17_{\pm 0.30}$	$21.25_{\pm 0.43}$	$18.43_{\pm 0.46}$	25.88
ToDi+Indep-2D	$26.70_{\pm 0.50}$	$36.30_{\pm 0.30}$	$33.17_{\pm 0.10}$	$20.17_{\pm 0.22}$	$19.21_{\pm 0.20}$	27.11	$20.83_{\pm 0.20}$	$32.47_{\pm 0.30}$	$35.37_{\pm 0.30}$	$21.00_{\pm 0.43}$	$18.47_{\pm 0.46}$	25.63
ToDi+JSVD	$\mathbf{27.70}_{\pm 0.66}$	$\mathbf{37.14}_{\pm 0.15}$	$34.11_{\pm 0.06}$	$20.66_{\pm 0.24}$	$\mathbf{20.13}_{\pm 0.18}$	**27.95**	$21.59_{\pm 0.27}$	$\mathbf{34.20}_{\pm 0.46}$	$37.10_{\pm 0.12}$	$22.34_{\pm 0.47}$	$\mathbf{19.64}_{\pm 0.37}$	26.97

positions based on learning difficulty. (iii) Decoupled 2D selection: Indep-2D, which forms an axis-factorized two-dimensional support by taking the Cartesian product of AdaKD-selected positions and Top-200 vocabulary candidates.

5.2 Overall Results

General Instruction-Following in LLMs. Tables 1 and 2 report the results on five instruction-following benchmarks. Under the Qwen2.5-7B→Qwen2.5-3B setting, JSVD consistently improves all four KD backbones and achieves the strongest average performance when combined with competitive objectives such as RKL and ToDi. The gains appear across both standard KL objectives and adaptive distillation objectives, suggesting that JSVD is not merely compensating for a specific divergence formulation. Instead, its advantage comes from allocating supervision more precisely over the sequence–vocabulary space.

The comparison with selective baselines further clarifies this point. Top-200 reduces vocabulary cost through a uniform teacher-side truncation rule, but ignores position-specific student errors. AdaKD selects informative sequence positions, yet still retains redundant vocabulary supervision at selected positions. Indep-2D sparsifies both sequence and vocabulary dimensions independently, but remains below JSVD across all backbones. This indicates that the key benefit is not simply using a sparse two-dimensional mask, but preserving the coupling between where the student makes an error and

Table 3: Results on mathematical reasoning and code generation benchmarks across different teacher–student settings. GSM8K and MATH are evaluated using exact-match accuracy, while HumanEval and MBPP are evaluated using Pass@1. Results on individual benchmarks are reported as mean ± standard deviation over five random seeds.

	Qwen2.5-7B → Qwen2.5-3B					
	Math			Code		
Method	**GSM8K**	**MATH**	**Avg.**	**HumanEval**	**MBPP**	**Avg.**
Teacher	86.31 ± 0.17	48.10 ± 0.24	67.21	67.94 ± 0.16	72.54 ± 0.10	70.24
Student	78.32 ± 0.26	39.44 ± 0.19	58.88	36.71 ± 0.20	61.73 ± 0.11	49.22
AKL	79.50 ± 0.14	43.85 ± 0.08	61.68	45.43 ± 0.11	62.97 ± 0.16	54.20
AKL+Top-200	78.89 ± 0.09	43.50 ± 0.13	61.20	44.30 ± 0.14	62.61 ± 0.05	53.46
AKL+AdaKD	80.30 ± 0.12	44.92 ± 0.14	62.61	47.13 ± 0.10	63.77 ± 0.12	55.45
AKL+Indep-2D	80.01 ± 0.09	44.30 ± 0.13	62.15	47.00 ± 0.14	63.45 ± 0.10	55.23
AKL+JSVD	81.00 ± 0.06	45.42 ± 0.10	63.21	47.40 ± 0.05	64.24 ± 0.14	55.82
ToDi	79.72 ± 0.20	44.61 ± 0.18	62.17	45.33 ± 0.11	63.40 ± 0.09	54.37
ToDi+Top-200	79.27 ± 0.03	44.13 ± 0.10	61.70	44.67 ± 0.14	62.80 ± 0.12	53.74
ToDi+AdaKD	80.67 ± 0.11	45.73 ± 0.09	63.20	47.10 ± 0.12	64.17 ± 0.11	55.64
ToDi+Indep-2D	80.60 ± 0.12	45.40 ± 0.14	63.00	46.70 ± 0.10	63.99 ± 0.14	55.35
ToDi+JSVD	$\mathbf{81.40 \pm 0.07}$	$\mathbf{46.15 \pm 0.10}$	**63.78**	$\mathbf{47.66 \pm 0.13}$	$\mathbf{64.70 \pm 0.09}$	**56.18**

	OpenLLaMA2-7B → OpenLLaMA2-3B						Qwen2.5-7B → Qwen2.5-1.5B					
	Math			Code			Math			Code		
Method	**GSM8K**	**MATH**	**Avg.**	**HumanEval**	**MBPP**	**Avg.**	**GSM8K**	**MATH**	**Avg.**	**HumanEval**	**MBPP**	**Avg.**
Teacher	83.31 ± 0.10	47.28 ± 0.39	65.30	60.29 ± 0.30	65.30 ± 0.21	62.80	86.31 ± 0.17	48.10 ± 0.24	67.21	67.94 ± 0.16	72.54 ± 0.10	70.24
Student	75.92 ± 0.26	30.73 ± 0.19	53.33	34.71 ± 0.17	57.80 ± 0.26	46.26	70.57 ± 0.26	23.10 ± 0.19	46.84	28.94 ± 0.40	53.48 ± 0.29	41.21
AKL	77.63 ± 0.20	33.20 ± 0.17	55.42	40.86 ± 0.10	58.77 ± 0.23	49.82	71.82 ± 0.14	37.49 ± 0.10	54.66	35.79 ± 0.24	55.86 ± 0.14	45.83
AKL+Top-200	75.40 ± 0.14	29.91 ± 0.10	52.66	37.23 ± 0.39	57.43 ± 0.24	47.33	71.09 ± 0.23	33.60 ± 0.11	52.35	32.10 ± 0.30	53.67 ± 0.21	42.89
AKL+AdaKD	78.98 ± 0.06	34.56 ± 0.13	56.77	42.90 ± 0.11	59.24 ± 0.18	51.07	72.50 ± 0.14	38.69 ± 0.10	55.60	37.60 ± 0.27	56.09 ± 0.21	46.85
AKL+Indep-2D	78.14 ± 0.24	34.50 ± 0.30	56.32	41.36 ± 0.10	58.70 ± 0.23	50.03	72.16 ± 0.13	38.70 ± 0.20	55.43	36.28 ± 0.27	55.79 ± 0.21	46.04
AKL+JSVD	80.85 ± 0.08	36.20 ± 0.14	58.53	44.59 ± 0.10	61.10 ± 0.07	52.85	74.08 ± 0.13	39.94 ± 0.20	57.01	38.73 ± 0.27	57.23 ± 0.21	47.98
ToDi	78.55 ± 0.13	34.80 ± 0.30	56.68	42.94 ± 0.20	58.70 ± 0.21	50.82	74.21 ± 0.20	38.55 ± 0.19	56.38	40.84 ± 0.24	56.74 ± 0.14	48.79
ToDi+Top-200	75.83 ± 0.19	29.31 ± 0.14	52.57	40.13 ± 0.30	57.59 ± 0.20	48.86	73.04 ± 0.13	37.25 ± 0.31	55.15	39.26 ± 0.20	55.30 ± 0.26	47.28
ToDi+AdaKD	79.38 ± 0.20	35.50 ± 0.06	57.44	44.01 ± 0.10	59.96 ± 0.13	51.99	75.22 ± 0.28	39.45 ± 0.20	57.34	41.53 ± 0.19	57.39 ± 0.10	49.46
ToDi+Indep-2D	78.90 ± 0.22	35.14 ± 0.17	57.02	43.14 ± 0.32	59.43 ± 0.20	51.29	74.90 ± 0.21	38.45 ± 0.16	56.68	41.00 ± 0.19	56.92 ± 0.10	48.96
ToDi+JSVD	$\mathbf{81.87 \pm 0.16}$	$\mathbf{37.02 \pm 0.10}$	**59.45**	$\mathbf{45.01 \pm 0.15}$	$\mathbf{61.19 \pm 0.17}$	**53.10**	$\mathbf{75.76 \pm 0.13}$	$\mathbf{40.01 \pm 0.10}$	**57.89**	$\mathbf{41.97 \pm 0.08}$	$\mathbf{58.06 \pm 0.11}$	**50.02**

which vocabulary candidates remain unresolved. Thus, high-value distillation signals are better captured as localized (t, u) pairs than by marginal sequence or vocabulary scores.

Table 2 shows that the same trend holds under additional teacher–student settings, including Qwen2.5-7B→Qwen2.5-1.5B and OpenLLaMA2-7B→OpenLLaMA2-3B. Across different model families, compression ratios, and KD objectives, JSVD yields stable improvements over the corresponding backbones. These results demonstrate that adaptive pair-level selection generalizes beyond a single architecture or model scale, providing a robust strategy for improving instruction-following distillation.

Results on Mathematical Reasoning and Code Generation. Table 3 evaluates JSVD on mathematical reasoning and code generation across three teacher–student settings. Under Qwen2.5-7B→Qwen2.5-3B, JSVD improves AKL from 61.68 to 63.21 in average exact-match accuracy on GSM8K/MATH, and from 54.20 to 55.82 in average Pass@1 on HumanEval/MBPP. Similar improvements are observed with ToDi, increasing the corresponding averages from 62.17 to 63.78 and from 54.37 to 56.18. The same trend persists across different model families and compression scales: under OpenLLaMA2-7B→OpenLLaMA2-3B, AKL+JSVD improves the math and code averages from 55.42 to 58.53 and from 49.82 to 52.85, respectively; under Qwen2.5-7B→Qwen2.5-1.5B, JSVD likewise yields consistent gains for both AKL and ToDi. Compared with selective baselines, Top-200 often degrades the backbone, while AdaKD and Indep-2D generally improve performance but remain below JSVD. These results show that preserving fine-grained sequence–vocabulary coupling generalizes beyond instruction-following to mathematical reasoning and code generation, and remains effective across model families and student scales.

5.3 Ablation Studies

Ablation Study of the Bidirectional Top-K Selection Module. Unless otherwise specified, all ablations use the FKL backbone under Qwen2.5-7B→Qwen2.5-3B, following the same model and evaluation setup as FKL+JSVD in Table 1. We compare teacher-only Top-B selection with bidirectional selection BTS(U_B) on DollyEval, S-NI, and UnNI. Teacher-only Top-B retains the teacher's top-B candidates, whereas BTS(U_B)

Table 4: Ablation of bidirectional Top-K candidate selection under the FKL backbone.

Method	**Dolly**	**S-NI**	**UnNI**	**Avg.**
+Top-50	22.40	36.61	39.96	32.99
+BTS(U50)	22.87	37.73	40.88	33.83
+Top-100	23.56	37.93	42.03	34.51
+BTS(U100)	25.03	38.79	42.20	35.34
+Top-200	24.45	39.00	42.28	35.24
+BTS(U200)	**26.65**	**40.45**	**43.69**	**36.93**

constructs the union of the teacher's and student's top-$B/2$ candidates, with at most B distinct candidates. As shown in Table 4, BTS consistently outperforms teacher-only selection across the tested budgets. At $B = 200$, BTS achieves an average ROUGE-L of 36.93, compared with 35.24 for teacher-only selection. These results suggest that incorporating student-side candidates provides additional informative supervision. We adopt BTS(U_{200}) in the main experiments, corresponding to the FKL+JSVD configuration in Table 1.

Analysis of Adaptive Difficulty Rate. To evaluate the Adaptive Difficulty Rate (ADR) mechanism, which dynamically adjusts the focus ratio in JSVD, we replace its adaptive update with fixed ratios $r \in \{0.4, 0.6, 0.8\}$, while keeping the FKL backbone, bidirectional candidate construction (BTS(U_{200})), pair ranking, and optimization settings unchanged. Here, BTS(U_{200}) uses the union of the teacher's and student's top-100 candidates. As shown in Table 5, ADR achieves the highest average ROUGE-L of 36.93, outperforming the strongest fixed ratio $r = 0.8$ (35.88) by 1.05 points. Reducing the fixed ratio to $r = 0.6$ and $r = 0.4$ decreases performance by 3.08 and 4.15 points, respectively, relative to ADR. These results demonstrate that dynamically adjusting the focus ratio better balances supervision coverage and selective focusing than fixed-ratio strategies.

Table 5: Ablation of the adaptive focus-ratio mechanism under FKL.

Method	Dolly	S-NI	UnNI	Avg.
$r = 0.4$	21.66	35.84	40.85	32.78
$r = 0.6$	23.54	36.13	41.88	33.85
$r = 0.8$	25.68	39.36	42.60	35.88
ADR	**26.65**	**40.45**	**43.69**	**36.93**

Training Stability and Convergence. We further analyze the training dynamics under the FKL backbone in Figure 3. Figure 3(a) shows that JSVD achieves the fastest loss reduction and the lowest held-out distillation loss, outperforming Top-200, AdaKD, and Indep-2D in optimization efficiency. Although Indep-2D also sparsifies the sequence–vocabulary plane, its slower and higher convergence suggests that preserving pair-level coupling is more effective than independently selecting the two dimensions. Figure 3(b) shows that this advantage translates into downstream performance, with JSVD reaching higher ROUGE-L earlier and maintaining the best validation performance for most of training. Finally, Figure 3(c) shows that the focus ratio r_s gradually decreases, indicating a shift from broad supervision to concentrated correction of unresolved high-discrepancy pairs. These results show that JSVD improves both convergence and generalization by dynamically allocating supervision to informative sequence–vocabulary pairs.

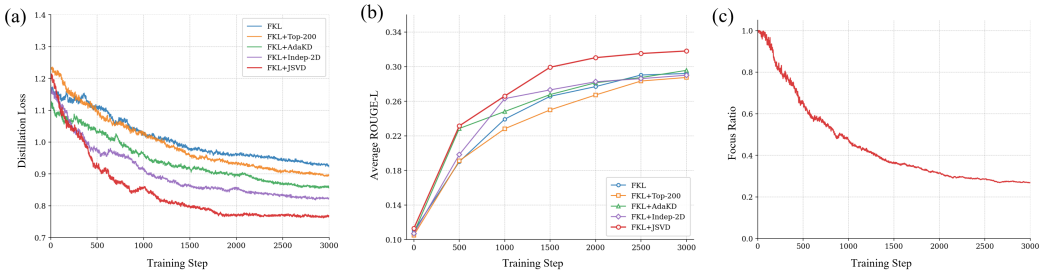


Figure 3: Training dynamics under the FKL backbone: (a) held-out distillation loss, (b) validation ROUGE-L, and (c) JSVD focus ratio r_s.

Table 6: End-to-end training efficiency and pair-budget-matched comparison under the FKL backbone. Profiling includes teacher- and student-side computation and selection overhead. Efficiency results are reported as mean ± standard deviation over three profiling runs.

Method	Cand./active pos.	Pairs/token	Active pos.	E2E ms/step ↓	Tokens/s ($\times 10^3$) ↑	Memory (GB) ↓	PFLOPs/step ↓	Avg. ROUGE-L ↑
Full KD	151,936	151,936	100%	1921 ± 10	63.4 ± 0.7	67.9 ± 0.2	4.26 ± 0.03	29.50
Top-200	200	200	100%	1736 ± 10	70.2 ± 0.8	63.4 ± 0.2	4.07 ± 0.04	29.21
AdaKD	151,936	112,433	74%	1849 ± 17	65.9 ± 0.5	65.7 ± 0.3	4.18 ± 0.05	29.81
Indep-2D	200	144	72%	1687 ± 20	72.3 ± 0.8	62.9 ± 0.3	3.93 ± 0.02	29.76
Indep-2D (matched)	168	121	72%	1603 ± 20	75.9 ± 0.9	62.9 ± 0.3	3.93 ± 0.01	28.60
JSVD	**161**	**121**	**77%**	$\mathbf{1590 \pm 15}$	$\mathbf{76.6 \pm 0.7}$	$\mathbf{61.8 \pm 0.2}$	$\mathbf{3.81 \pm 0.03}$	**31.29**

End-to-End Efficiency and Budget-Matched Comparison. Table 6 evaluates the end-to-end training efficiency of JSVD under the FKL backbone, including teacher and student computation, candidate construction and selection, backpropagation, and optimizer updates. All methods use

identical batches, data order, padding patterns, and gradient accumulation settings, with efficiency measurements averaged over three profiling runs.

For JSVD, the default BTS(U_{200}) constructs the union of the teacher's and student's top-100 candidates. Due to candidate overlap, the actual support contains at most 200 distinct vocabulary candidates, with an observed average of 161 candidates per active sequence position.

Compared with Full KD, JSVD reduces end-to-end step time from 1921 to 1590 ms, achieving a $1.21\times$ speedup while reducing peak memory from 67.9 to 61.8 GB and total computation from 4.26 to 3.81 PFLOPs/step. Meanwhile, JSVD improves average ROUGE-L from 29.50 to 31.29, demonstrating simultaneous efficiency and effectiveness gains.

To distinguish the benefit of joint pair-level selection from differences in supervision budget, we further compare JSVD with Indep-2D under a matched average budget of 121 selected pairs per token. The matched Indep-2D baseline adjusts its vocabulary budget while retaining the original position selector. Under this setting, both methods exhibit comparable end-to-end step times (1590 ± 15 vs. 1603 ± 20 ms), yet JSVD achieves a 2.69-point improvement in average ROUGE-L, together with lower peak memory and total computation. These results suggest that JSVD's effectiveness cannot be explained solely by the number of selected pairs, highlighting the benefit of joint sequence–vocabulary selection over independent axis-wise selection.

6 Conclusion and Limitations

We present JSVD, a plug-and-play selective distillation framework that operates directly in the sequence–vocabulary space. By focusing supervision on unresolved high-value pairs, JSVD consistently improves multiple KD backbones across instruction-following, mathematical reasoning, and code generation tasks (Tables 1, 2, 3). Compared with single-axis selection methods such as Top-200 and AdaKD, as well as the dimension-decoupled Indep-2D baseline, JSVD preserves pair-level coupling and achieves stronger performance while reducing distillation overhead and accelerating convergence (Figure 3, Table **??**). These results suggest that adaptive pair-level selection is a general and architecture-agnostic strategy for allocating distillation supervision more effectively.

As a plug-and-play selective distillation method, JSVD does not introduce architectural constraints on the teacher or student model, nor does it depend on a specific logit-based KD objective. Its main practical limitation is that pair-level selection is restricted to the candidate support constructed before scoring; therefore, useful low-probability but task-relevant tokens may be missed if they are excluded from the bidirectional top-k support. Future work may address this issue through more adaptive support construction, including mass-aware expansion, uncertainty-guided inclusion, or dynamic candidate budgets. By reducing the cost of dense logit distillation, JSVD can facilitate more efficient compression and deployment of LLMs under limited computational resources. We do not foresee additional societal risks beyond those generally associated with the use and deployment of LLMs.

Acknowledgements and Disclosures

Acknowledgements. We appreciate the anonymous reviewers for their valuable insights and helpful comments.

Funding. This work was supported in part by the Ministry of Education Top-notch Student Training Program in Basic Disciplines 2.0 Research Topics (No. 20252012), the National Key R&D Program of China (No. 2023YFB3107500), the National Natural Science Foundation of China (Nos. 62502362, 62606380, 92467201, and 62220106004), the Open Project of Shandong Provincial Key Laboratory of Independent and Reliable Computing Technology and Equipment (No. KT25500400-lab), the Natural Science Basic Research Program of Shaanxi Province (Nos. 2025JC-YBQN-869, 2026JC-YBQN-0807, and 2026JC-YXQN-230), the China Postdoctoral Science Foundation (Nos. 2025M771506 and 2025M781513), the Xidian University Specially Funded Project for Interdisciplinary Exploration (No. TZJHF202502), and the Fundamental Research Funds for the Central Universities (Nos. QTZX26050, QTZX26022, QTZX26041, and QTZX26118).

Competing Interests. The authors declare no competing interests.

References

[1] Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Ross Mitchell. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms, 2026. URL https://arxiv.org/abs/2510.27246.

[2] Dilip Arumugam and Thomas L. Griffiths. Toward efficient exploration by large language model agents, 2026. URL https://arxiv.org/abs/2504.20997.

[3] Shiqi Yan, Yubo Chen, Ruiqi Zhou, Zhengxi Yao, Shuai Chen, Tianyi Zhang, Shijie Zhang, Wei Qiang Zhang, Yongfeng Huang, Haixin Duan, and Yunqi Zhang. Explore-on-graph: Incentivizing autonomous exploration of large language models on knowledge graphs with path-refined reward modeling, 2026. URL https://arxiv.org/abs/2602.21728.

[4] Haochen Luo, Ho Tin Ko, Jiandong Chen, David Sun, Yuan Zhang, and Chen Liu. Alphabench: Benchmarking large language models in formulaic alpha factor mining. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=d97Q8r7ZKZ.

[5] Erfan Jahanparast, Zhiqing Hong, and Serina Chang. What do large language models know about opinions? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kHVzEjThKE.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[7] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

[8] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models, 2024. URL https://arxiv.org/abs/2402.13116.

[9] Hao Peng, Xin Lv, Yushi Bai, Zijun Yao, Jiajie Zhang, Lei Hou, and Juanzi Li. Pre-training distillation for large language models: A design space exploration, 2024. URL https://arxiv.org/abs/2410.16215.

[10] Anshumann Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse logit sampling: Accelerating knowledge distillation in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18085–18108, 2025.

[11] Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. Selective knowledge distillation for neural machine translation. *arXiv preprint arXiv:2105.12967*, 2021.

[12] Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. Selectkd: Selective token-weighted knowledge distillation for llms. *arXiv preprint arXiv:2510.24021*, 2025.

[13] Xurong Xie, Zhucun Xue, Jiafu Wu, Jian Li, Yabiao Wang, Xiaobin Hu, Yong Liu, and Jiangning Zhang. Llm-oriented token-adaptive knowledge distillation. *arXiv preprint arXiv:2510.11615*, 2025.

[14] Xiaoyu Liu, Yun Zhang, Wei Li, Simiao Li, Xudong Huang, Hanting Chen, Yehui Tang, Jie Hu, Zhiwei Xiong, and Yunhe Wang. Multi-granularity semantic revision for large language model distillation. *arXiv preprint arXiv:2407.10068*, 2024.

[15] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4182–4191, 2020. doi: 10.1109/CVPR42600.2020.00424.

[16] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022.

[17] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

[18] Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.

[19] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 14852–14882, 2023.

[20] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, 2023.

[21] Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5737–5755, 2025.

[22] Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. Todi: Token-wise distillation via fine-grained divergence control. *arXiv preprint arXiv:2505.16297*, 2025.

[23] Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. Abkd: Pursuing a proper allocation of the probability mass in knowledge distillation via $alpha$-$beta$-divergence. *arXiv preprint arXiv:2505.04560*, 2025.

[24] Yuezhou Hu, Jiaxin Guo, Xinyu Feng, and Tuo Zhao. Adaspec: Selective knowledge distillation for efficient speculative decoders, 2025. URL `https://arxiv.org/abs/2510.19779`.

[25] Mrigank Raman, Pranav Mani, Davis Liang, and Zachary Lipton. For distillation, tokens are not all you need. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

[26] Almog Tavor, Itay Ebenspanger, Neil Cnaan, and Mor Geva. Rethinking selective knowledge distillation. *arXiv preprint arXiv:2602.01395*, 2026.

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

[28] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL `https://arxiv.org/abs/2501.00656`.

[29] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL `https://arxiv.org/abs/2411.15124`.

[30] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instructiontuned llm. 2023.

[31] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508, 2023.

[32] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[33] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*, 2022.

[34] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, 2023.

[35] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

[36] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[37] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[38] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023.

[39] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[40] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[41] Xurong Xie, Zhucun Xue, Jiafu Wu, Jian Li, Yabiao Wang, Xiaobin Hu, Yong Liu, and Jiangning Zhang. Llm-oriented token-adaptive knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34070–34078, 2026.

A Broader Impacts

This work studies selective knowledge distillation for large language models, with the goal of improving the efficiency and effectiveness of logit-based distillation. A positive societal impact of JSVD is that it can reduce the computational overhead of transferring knowledge from large teacher models to smaller student models. By focusing supervision on informative sequence–vocabulary pairs, JSVD may help lower the cost, memory usage, and energy consumption associated with LLM compression. This can make model compression more accessible to researchers and practitioners with limited computational resources, and may support deployment in resource-constrained environments.

JSVD does not introduce new model capabilities, datasets, or deployment scenarios beyond those already associated with LLM distillation. However, as with other methods that make LLM compression more efficient, it may lower the cost of adapting or deploying generative models. If applied to

unsafe teacher models, uncurated training data, or poorly governed deployment settings, the distilled student may inherit undesirable behaviors such as hallucination, biased responses, toxic generation, or privacy leakage. These risks are not specific to JSVD, but reflect broader concerns associated with the use and deployment of large language models.

We therefore encourage practitioners to apply JSVD with appropriately curated training data, reliable teacher models, and task-specific safety evaluation. For distilled models intended for public release, additional safeguards such as bias evaluation, harmful-content filtering, red-teaming, license compliance, and controlled release mechanisms may be necessary.

B Proofs for Discrepancy-Based Pair Scoring

This appendix provides the proofs for the claims in Section 4.4. We first relate the raw full-support discrepancy used for scoring to the renormalized discrepancy used in the restricted-support KL loss. We then prove Proposition 1, which establishes the local optimization meaning of discrepancy-based scoring, and Proposition 2, which shows why axis-wise selection cannot recover pair-level structure.

B.1 Support-Renormalization Consistency

Let $p_t^{\mathcal{V}}$ and $q_t^{\mathcal{V}}$ be the teacher and student distributions over the full vocabulary $\mathcal{V}$. For a selected support $A_t \subseteq \mathcal{V}$, define

$$P_A = \sum_{v \in A_t} p_t^{\mathcal{V}}(v), \qquad Q_A = \sum_{v \in A_t} q_t^{\mathcal{V}}(v), \qquad m_A = \min\{P_A, Q_A\},$$

and the restricted normalized distributions

$$p_t^A(u) = \frac{p_t^{\mathcal{V}}(u)}{P_A}, \qquad q_t^A(u) = \frac{q_t^{\mathcal{V}}(u)}{Q_A}, \qquad u \in A_t.$$

For any $u \in A_t$ and $m_A > 0$, we have

$$\left|p_t^A(u) - q_t^A(u)\right| \le \frac{\left|p_t^{\mathcal{V}}(u) - q_t^{\mathcal{V}}(u)\right| + |P_A - Q_A|}{m_A}, \tag{15}$$

and

$$\left|p_t^{\mathcal{V}}(u) - q_t^{\mathcal{V}}(u)\right| \le \left|p_t^A(u) - q_t^A(u)\right| + |P_A - Q_A|. \tag{16}$$

Proof. For Eq. (15),

$$\begin{aligned} p_t^A(u) - q_t^A(u) &= \frac{p_t^{\mathcal{V}}(u)}{P_A} - \frac{q_t^{\mathcal{V}}(u)}{Q_A} \\ &= \frac{Q_A\big(p_t^{\mathcal{V}}(u) - q_t^{\mathcal{V}}(u)\big) + q_t^{\mathcal{V}}(u)(Q_A - P_A)}{P_A Q_A}. \end{aligned}$$

Taking absolute values and using $q_t^{\mathcal{V}}(u) \le Q_A$ gives

$$\left|p_t^A(u) - q_t^A(u)\right| \le \frac{\left|p_t^{\mathcal{V}}(u) - q_t^{\mathcal{V}}(u)\right| + |P_A - Q_A|}{P_A} \le \frac{\left|p_t^{\mathcal{V}}(u) - q_t^{\mathcal{V}}(u)\right| + |P_A - Q_A|}{m_A}.$$

For Eq. (16), since $p_t^{\mathcal{V}}(u) = P_A p_t^A(u)$ and $q_t^{\mathcal{V}}(u) = Q_A q_t^A(u)$,

$$\begin{aligned} \left|p_t^{\mathcal{V}}(u) - q_t^{\mathcal{V}}(u)\right| &= \left|P_A p_t^A(u) - Q_A q_t^A(u)\right| \\ &= \left|P_A\big(p_t^A(u) - q_t^A(u)\big) + (P_A - Q_A) q_t^A(u)\right| \\ &\le P_A \left|p_t^A(u) - q_t^A(u)\right| + |P_A - Q_A| q_t^A(u) \\ &\le \left|p_t^A(u) - q_t^A(u)\right| + |P_A - Q_A|. \end{aligned}$$

□

Moreover,

$$|P_A - Q_A| = \left| \sum_{v \in A_t} \left(p_t^{\mathcal{V}}(v) - q_t^{\mathcal{V}}(v) \right) \right| \leq \sum_{v \in A_t} \left| p_t^{\mathcal{V}}(v) - q_t^{\mathcal{V}}(v) \right|.$$

Therefore, when A_t has non-negligible probability mass and the aggregate teacher–student mass mismatch on A_t is small, the raw discrepancy and the restricted normalized discrepancy are locally order-equivalent.

B.2 Proof of Proposition 1

Let $p_t, q_t \in \Delta(A_t)$ be normalized teacher and student distributions on a fixed support A_t. Define

$$\Delta_t = q_t - p_t, \qquad \mathcal{T}_t = \{v \in \mathbb{R}^{|A_t|} : \mathbf{1}^\top v = 0\}.$$

Since both distributions lie on the same simplex, $\Delta_t \in \mathcal{T}_t$.

Assume that $D(p_t, q_t)$ is twice continuously differentiable near $q_t = p_t$, minimized at $q_t = p_t$, and has no nonzero flat direction on $\mathcal{T}_t$. Let P_t be the orthogonal projection onto $\mathcal{T}_t$. The constrained first-order optimality condition gives

$$P_t \nabla_q D(p_t, q_t)\big|_{q_t = p_t} = 0.$$

For sufficiently small $\|\Delta_t\|_2$, the path $p_t + \alpha \Delta_t$, $\alpha \in [0, 1]$, remains in the relative interior of the simplex. Taylor expansion along $\mathcal{T}_t$ yields

$$D(p_t, q_t) = \frac{1}{2} \Delta_t^\top \nabla_q^2 D(p_t, q_t)\big|_{q_t = p_t} \Delta_t + o(\|\Delta_t\|_2^2).$$

Define

$$G_{p_t} = P_t \left(\nabla_q^2 D(p_t, q_t)\big|_{q_t = p_t} \right) P_t.$$

Then

$$D(p_t, q_t) = \frac{1}{2} \Delta_t^\top G_{p_t} \Delta_t + o(\|\Delta_t\|_2^2).$$

The non-degeneracy assumption ensures that G_{p_t} is positive definite on $\mathcal{T}_t$.

For the gradient, Taylor expansion of the projected q-gradient gives

$$P_t \nabla_q D(p_t, q_t) = G_{p_t} \Delta_t + o(\|\Delta_t\|_2).$$

By the chain rule,

$$\nabla_{z_t} D(p_t, q_t) = J(q_t) \nabla_{q_t} D(p_t, q_t),$$

where $J(q_t) = \operatorname{diag}(q_t) - q_t q_t^\top$ is the softmax Jacobian. Any component of $\nabla_{q_t} D$ parallel to $\mathbf{1}$ is annihilated by $J(q_t)$, since $J(q_t)\mathbf{1} = 0$. Thus,

$$\nabla_{z_t} D(p_t, q_t) = J(q_t) \left(G_{p_t} \Delta_t + o(\|\Delta_t\|_2) \right).$$

Since $J(q_t) = J(p_t) + O(\|\Delta_t\|_2)$, we obtain

$$\nabla_{z_t} D(p_t, q_t) = J(p_t) G_{p_t} \Delta_t + o(\|\Delta_t\|_2).$$

This proves Proposition 1.

B.3 KL Instantiations and Adaptive Variants

For FKL,

$$D_{\mathrm{FKL}}(p_t, q_t) = \sum_{u \in A_t} p_t(u) \log \frac{p_t(u)}{q_t(u)}.$$

The Hessian with respect to q_t is

$$\nabla_q^2 D_{\mathrm{FKL}}(p_t, q_t) = \operatorname{diag}\left(\frac{p_t(u)}{q_t(u)^2} \right)_{u \in A_t}.$$

At $q_t = p_t$,

$$\nabla_q^2 D_{\mathrm{FKL}}(p_t, q_t)\big|_{q_t=p_t} = \mathrm{diag}\left(\frac{1}{p_t(u)}\right)_{u\in A_t}.$$

Hence, on the tangent space,

$$D_{\mathrm{FKL}}(p_t, q_t) = \frac{1}{2}\sum_{u\in A_t}\frac{\Delta_t(u)^2}{p_t(u)} + o(\|\Delta_t\|_2^2). \tag{17}$$

The FKL logit gradient has the exact form

$$\frac{\partial D_{\mathrm{FKL}}(p_t, q_t)}{\partial z_t(u)} = q_t(u) - p_t(u).$$

Indeed,

$$\begin{aligned}\frac{\partial D_{\mathrm{FKL}}}{\partial z_t(u)} &= -\sum_{v\in A_t} p_t(v)\frac{\partial \log q_t(v)}{\partial z_t(u)}\\ &= -\sum_{v\in A_t} p_t(v)\left(\delta_{uv} - q_t(u)\right)\\ &= -p_t(u) + q_t(u)\sum_{v\in A_t} p_t(v)\\ &= q_t(u) - p_t(u).\end{aligned}$$

Therefore, under FKL,

$$\left|\frac{\partial D_{\mathrm{FKL}}(p_t, q_t)}{\partial z_t(u)}\right| = |\Delta_t(u)|.$$

For RKL,

$$D_{\mathrm{RKL}}(p_t, q_t) = \sum_{u\in A_t} q_t(u)\log\frac{q_t(u)}{p_t(u)}.$$

Its Hessian with respect to q_t is

$$\nabla_q^2 D_{\mathrm{RKL}}(p_t, q_t) = \mathrm{diag}\left(\frac{1}{q_t(u)}\right)_{u\in A_t},$$

and therefore at $q_t = p_t$,

$$D_{\mathrm{RKL}}(p_t, q_t) = \frac{1}{2}\sum_{u\in A_t}\frac{\Delta_t(u)^2}{p_t(u)} + o(\|\Delta_t\|_2^2). \tag{18}$$

The RKL logit gradient is

$$\frac{\partial D_{\mathrm{RKL}}(p_t, q_t)}{\partial z_t(u)} = q_t(u)\left(\log\frac{q_t(u)}{p_t(u)} - D_{\mathrm{RKL}}(p_t, q_t)\right).$$

If $p_t(u) \geq p_{\min} > 0$ for all $u \in A_t$, then as $q_t \to p_t$,

$$\log\frac{q_t(u)}{p_t(u)} = \frac{\Delta_t(u)}{p_t(u)} + O\left(\frac{\Delta_t(u)^2}{p_{\min}^2}\right),$$

and

$$D_{\mathrm{RKL}}(p_t, q_t) = O\left(\frac{\|\Delta_t\|_2^2}{p_{\min}}\right).$$

Substituting these expansions gives

$$\frac{\partial D_{\mathrm{RKL}}(p_t, q_t)}{\partial z_t(u)} = \Delta_t(u) + O\left(\frac{\|\Delta_t\|_2^2}{p_{\min}}\right).$$

Thus, discrepancy is the leading-order coordinate-wise logit-gradient signal for RKL.

The same argument extends to smooth adaptive KL variants. Consider a local mixture

$$D_{\mathrm{mix}}(p_t, q_t) = a(p_t, q_t) D_{\mathrm{FKL}}(p_t, q_t) + b(p_t, q_t) D_{\mathrm{RKL}}(p_t, q_t),$$

where a and b are continuously differentiable and locally bounded near $q_t = p_t$, and

$$\gamma_t = a(p_t, p_t) + b(p_t, p_t) > 0.$$

Because both KL terms vanish at $q_t = p_t$, derivatives of the weights only contribute higher-order terms. Combining the FKL and RKL expansions yields

$$D_{\mathrm{mix}}(p_t, q_t) = \frac{\gamma_t}{2} \sum_{u \in A_t} \frac{\Delta_t(u)^2}{p_t(u)} + o(\|\Delta_t\|_2^2),$$

and

$$\frac{\partial D_{\mathrm{mix}}(p_t, q_t)}{\partial z_t(u)} = \gamma_t \Delta_t(u) + O(\|\Delta_t\|_2^2).$$

Therefore, for smooth KL-based adaptive objectives with positive bounded local weights, discrepancy remains a leading-order coordinate-wise gradient proxy up to a positive local scaling factor.

B.4 Proof of Proposition 2

Let

$$H \in \mathbb{R}_+^{L \times |\mathcal{V}|}$$

denote a sequence–vocabulary discrepancy matrix, where $H_{t,u}$ is the discrepancy score of pair (t, u). For a pair support $\Omega \subseteq [L] \times \mathcal{V}$ with $|\Omega| = K$, define its selected discrepancy mass as

$$F_H(\Omega) = \sum_{(t,u) \in \Omega} H_{t,u}.$$

The optimal top-K pair support is therefore

$$\Omega_K^\star(H) \in \arg\max_{|\Omega| = K} F_H(\Omega).$$

Define the row and column aggregate scores by

$$r(H)_t = \sum_u H_{t,u}, \qquad c(H)_u = \sum_t H_{t,u}.$$

A marginal-based selector is any mapping of the form

$$\Phi_K(H) = \psi_K\big(r(H), c(H)\big),$$

which means that it can only use sequence-wise and vocabulary-wise aggregate scores, but not the full pair-level matrix.

It suffices to construct two discrepancy matrices with the same marginals but different optimal top-K supports. Consider the 2×2 case with $K = 2$. Let

$$\mathbf{J} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

and for any $0 < \epsilon < 1/2$, define

$$H^+ = \mathbf{J} + \epsilon B, \qquad H^- = \mathbf{J} - \epsilon B.$$

Explicitly,

$$H^+ = \begin{pmatrix} \frac{1}{2} + \epsilon & \frac{1}{2} - \epsilon \\ \frac{1}{2} - \epsilon & \frac{1}{2} + \epsilon \end{pmatrix}, \qquad H^- = \begin{pmatrix} \frac{1}{2} - \epsilon & \frac{1}{2} + \epsilon \\ \frac{1}{2} + \epsilon & \frac{1}{2} - \epsilon \end{pmatrix}.$$

Both matrices are strictly positive. Moreover, since every row and every column of B sums to zero, we have

$$r(H^+) = r(H^-) = (1, 1), \qquad c(H^+) = c(H^-) = (1, 1).$$

Thus, any marginal-based selector must output the same support for both matrices:

$$\Phi_2(H^+) = \Phi_2(H^-) = \Omega.$$

Let

$$D = \{(1,1),(2,2)\}, \qquad O = \{(1,2),(2,1)\}$$

denote the diagonal and off-diagonal supports. For H^+, the optimal top-2 support is D, with

$$F_{H^+}(D) = 1 + 2\epsilon.$$

For H^-, the optimal top-2 support is O, with

$$F_{H^-}(O) = 1 + 2\epsilon.$$

Now let

$$m = |\Omega \cap D|$$

be the number of diagonal entries selected by the marginal selector. Since $|\Omega| = 2$, we have $m \in \{0,1,2\}$. The selected masses on the two matrices are

$$F_{H^+}(\Omega) = 1 + \epsilon(2m - 2),$$

and

$$F_{H^-}(\Omega) = 1 + \epsilon(2 - 2m).$$

Therefore, the optimality gaps are

$$F_{H^+}(D) - F_{H^+}(\Omega) = 2\epsilon(2 - m),$$

and

$$F_{H^-}(O) - F_{H^-}(\Omega) = 2\epsilon m.$$

For the same support Ω to be optimal on both H^+ and H^-, we would need

$$2\epsilon(2 - m) = 0 \quad \text{and} \quad 2\epsilon m = 0,$$

which requires simultaneously $m = 2$ and $m = 0$, a contradiction. Equivalently,

$$\max\{F_{H^+}(D) - F_{H^+}(\Omega), F_{H^-}(O) - F_{H^-}(\Omega)\} \geq 2\epsilon.$$

Thus, for at least one of the two matrices, any selector based only on row and column marginals is strictly suboptimal.

This proves that sequence-wise and vocabulary-wise aggregates do not determine the optimal pair-level support. Consequently, high-value distillation signals cannot in general be recovered from independent axis-wise scores, and must instead be selected directly over sequence–vocabulary pairs (t, u).

The previous construction also recovers the intuitive binary example as a special case. Taking the boundary value $\epsilon = 1/2$ gives

$$H^+ = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad H^- = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Both matrices have identical sequence-wise and vocabulary-wise aggregates, but their optimal top-2 supports are respectively diagonal and off-diagonal. Thus, marginal scores cannot recover localized pair-level structure, whereas JSVD selects directly over sequence–vocabulary pairs (t, u).

C Default hyperparameter settings.

We report the default hyperparameter settings used in all experiments.

Table 7: Default hyperparameter settings.

Parameters	Values
β	0.97
α	0.05
Effective Batch Size	128
Dropout (p)	0.0
Sequence Length	1024
Train Iterations	3000
Learning Rate	3×10^{-4}
Schedule	Cosine
LR Decay Iterations	100%
Warmup Steps	10%
Min LR	3×10^{-5}
Gradient Clipping	1.0

D Training Algorithm

Algorithm 1 Training Process of JSVD

Require: Teacher p, student q_θ, dataset $\mathcal{D}$, top-K, iterations S, initial ratio r_0, EMA β, adaptation rate α, constant ϵ, KD weight λ, learning rate η, divergence $D^{(t,u)}(\cdot,\cdot)$

Ensure: Updated student q_θ

Initialize $r_0 \leftarrow 1.0$, $\bar{D}_0^{\text{ref}} \leftarrow 0$
for $s = 1$ to S **do**
  Sample mini-batch $\mathcal{B}_s \subset \mathcal{D}$
  for each $x \in \mathcal{B}_s$ **do**
    for $t = 1$ to L **do**
      $\mathcal{I}_t^p \leftarrow \text{TopK}(p_t), \mathcal{I}_t^q \leftarrow \text{TopK}(q_{\theta_s,t})$
      $\mathcal{V}_t^s \leftarrow \mathcal{I}_t^p \cup \mathcal{I}_t^q$
      $H_s(t,u) \leftarrow |p_t(u) - q_{\theta_s,t}(u)|,\ \forall u \in \mathcal{V}_t^s$
    end for
    $C_s(x) \leftarrow \{(t,u) : t \in [1,L],\ u \in \mathcal{V}_t^s\}$
  end for
  $\bar{D}_s \leftarrow \frac{1}{|\mathcal{B}_s|} \sum_{x \in \mathcal{B}_s} \frac{1}{|C_s(x)|} \sum_{(t,u) \in C_s(x)} H_s(t,u)$
  $e_s \leftarrow \frac{\bar{D}_s - \bar{D}_{s-1}^{\text{ref}}}{\bar{D}_{s-1}^{\text{ref}} + \epsilon}$
  $r_s \leftarrow \text{clip}_{[0,1]}(r_{s-1} + \alpha e_s)$
  $\bar{D}_s^{\text{ref}} \leftarrow \beta \bar{D}_{s-1}^{\text{ref}} + (1-\beta)\bar{D}_s$
  for each $x \in \mathcal{B}_s$ **do**
    $\Omega_s(x) \leftarrow Top_{r_s}\{H_s(t,u) : (t,u) \in C_s(x)\}$
    $A_t(x) \leftarrow \{u : (t,u) \in \Omega_s(x)\}$
    $\bar{p}_t(u) \leftarrow \frac{p_t(u)}{\sum_{v \in A_t(x)} p_t(v)}$
    $\bar{q}_t(u) \leftarrow \frac{q_{\theta_s,t}(u)}{\sum_{v \in A_t(x)} q_{\theta_s,t}(v)}$
  end for
  $\mathcal{L}_{\text{JSVD},s} \leftarrow \frac{1}{|\mathcal{B}_s|} \sum_{x \in \mathcal{B}_s} \sum_{t : A_t(x) \neq \emptyset} \sum_{u \in A_t(x)} D^{(t,u)}(\bar{p}_t, \bar{q}_t)$
  $\mathcal{L}_{\text{CE},s} \leftarrow -\frac{1}{|\mathcal{B}_s|} \sum_{x \in \mathcal{B}_s} \sum_{t=1}^{L} \log q_{\theta_s,t}(y_t)$
  $\mathcal{L}_{\text{total},s} \leftarrow (1-\lambda)\mathcal{L}_{\text{CE},s} + \lambda \mathcal{L}_{\text{JSVD},s}$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total},s}$
end for

E Existing Assets and Licenses

This work uses publicly available models, datasets, and evaluation benchmarks. We do not introduce a new dataset or release a new pretrained model. All existing assets are used for research purposes, and we cite their original papers or repositories in the main text. We follow the corresponding licenses and terms of use of the original asset providers. Table 8 summarizes the main existing assets used in this paper.

Table 8: Existing models, datasets, and benchmarks used in this paper. We use these assets for research purposes and follow the licenses and terms specified by their original providers.

Asset	Use in this paper	License / Terms	Credit / Source
Qwen2.5-7B and Qwen2.5-1.5B	Teacher or student LLMs in distillation experiments	Apache-2.0, according to the official Qwen2.5 release information	Qwen Team
Qwen2.5-3B	Student LLM in the Qwen2.5-7B→Qwen2.5-3B setting	Qwen Research License, according to the official model card / license file	Qwen Team
OpenLLaMA2-7B and OpenLLaMA2-3B	Teacher and student LLMs in additional model-family experiments	Apache-2.0	OpenLM Research
Tulu / OLMo2 SFT mixture	Instruction-following training data	ODC-BY-1.0 for the released mixture; subset-specific licenses may apply, including non-commercial terms for some subsets	Allen Institute for AI and associated subset creators
MetaMathQA	Mathematical reasoning distillation data	Public research dataset derived from GSM8K and MATH training sets; we follow the terms of the original MetaMathQA release and its source datasets	MetaMath authors
WizardCoder Evol-Instruct data	Code-generation training data	Public research dataset; we follow the terms specified by the original release	WizardCoder authors
DollyEval / Databricks Dolly data	Instruction-following evaluation / related instruction data	CC-BY-SA-3.0 for Databricks Dolly-15k	Databricks
Super-NaturalInstructions (S-NI)	Instruction-following evaluation	Apache-2.0 for the instruction collection, with instance-level licenses inherited from source datasets	Allen Institute for AI and dataset contributors
Unnatural Instructions (UnNI)	Instruction-following evaluation	Public research benchmark; we follow the terms specified by the original release	Honovich et al.
Self-Instruct	Instruction-following evaluation	Public research benchmark; we follow the terms specified by the original release	Wang et al.
VicunaEval	Instruction-following evaluation	Apache-2.0 for the public Vicuna evaluation repository	LMSYS / Vicuna authors
GSM8K	Mathematical reasoning evaluation	MIT	OpenAI
MATH	Mathematical reasoning evaluation	MIT	Hendrycks et al.
HumanEval	Code-generation evaluation	MIT	OpenAI
MBPP	Code-generation evaluation	Apache-2.0, following the Google Research repository	Austin et al. / Google Research

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction clearly state the main claims of the paper: JSVD formulates selective logit-based distillation in the sequence–vocabulary space, selects informative sequence–vocabulary pairs via bidirectional candidate construction, discrepancy-based scoring, and adaptive focusing, and empirically improves performance and efficiency across multiple KD objectives, model pairs, and benchmarks. These claims are supported by the methodology, theoretical analysis, and experiments reported in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The conclusion discusses the main limitation of JSVD: pair-level selection is restricted to the bidirectional candidate support, which may exclude low-probability but task-relevant vocabulary tokens. It also outlines potential improvements through adaptive support expansion, uncertainty-guided inclusion, and dynamic candidate budgets.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: The paper includes theoretical results on discrepancy-based pair scoring and the limitations of axis-wise selection in the main text, with complete proofs provided in Appendix B. Appendix B states the assumptions required for support renormalization and local analysis and provides formal derivations supporting the theoretical results used by JSVD.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The paper specifies the teacher–student model pairs, training datasets, evaluation benchmarks, metrics, LoRA configuration, number of random seeds, and the main distillation objectives and baselines. It also describes the JSVD selection procedure, including bidirectional candidate construction, discrepancy-based pair scoring, and adaptive focusing, which are necessary to reproduce the main experimental comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The paper uses publicly available models, datasets, and benchmarks, and provides the main algorithmic and experimental details needed to reproduce the results. However, the implementation code and scripts are not publicly released at submission time to avoid potential de-anonymization risks during the double-blind review process, such as accidentally exposing author-identifying information through file paths, logging metadata, account information, or repository history. We plan to release an anonymized or de-identified implementation after the review process, subject to the conference policy.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The paper reports the training configuration, model pairs, LoRA hyperparameters, datasets, evaluation benchmarks, evaluation metrics, number of random seeds, and baseline methods. Additional experimental details are summarized in the appendix, which supports interpretation and reproduction of the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The paper reports results over five random seeds and presents the main experimental tables using mean and standard deviation. The tables highlight the best and second-best scores, and the use of multiple seeds provides variability estimates for the main empirical claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: All experiments in the paper are conducted under the same compute setting using eight NVIDIA H100 80GB GPUs.The paper reports the hardware configuration used for training and provides end-to-end profiling results, including step time, token throughput, peak GPU memory, and PFLOPs per step, to characterize the computational cost of JSVD and the compared baselines.Since the same compute workers are used across all main experiments, this information provides a consistent basis for reproducing the reported results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work studies selective knowledge distillation for LLMs using publicly available models, datasets, and benchmarks. It does not involve human-subject experiments, private data collection, or deployment in sensitive decision-making settings, and the research is presented as a method for improving the efficiency of model distillation.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper includes a Broader Impacts section in the appendix. It discusses positive impacts such as reducing the computational cost, memory usage, and energy consumption of LLM distillation, as well as potential negative impacts such as lowering the barrier to distilling or deploying models that may generate harmful, misleading, biased, or unsafe content.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

Justification: The paper does not introduce or release a new high-risk pretrained language model, image generator, or scraped dataset. It proposes a distillation algorithm and evaluates it using existing models and benchmark datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The paper uses publicly available models, datasets, and benchmarks, and cites the original sources in the main text. We additionally provide an appendix section summarizing the main existing assets, their roles in the experiments, and their licenses or terms of use. All assets are used for research purposes in accordance with the corresponding licenses and provider terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A].

Justification: The paper does not introduce a new dataset, benchmark, pretrained model, or other released asset. The contribution is a selective distillation algorithm evaluated on existing models and datasets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All evaluations are conducted on existing public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing, user studies, or other human-subject research, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: The paper's core method is designed for logit-based knowledge distillation of large language models. It explicitly describes the teacher and student LLMs, the autoregressive distillation setting, the model pairs used in experiments, and how teacher and student output distributions are used to construct JSVD supervision.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.