EVIDENCE-ADAPTIVE DUAL-SPACE DISTILLATION FOR LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-tokenizer knowledge distillation requires both alignment between heterogeneous output spaces and an appropriate allocation of supervision across them. We introduce **Evidence-Adaptive Knowledge Distillation (EAKD)**, which uses detached student and teacher predictive cross-entropies to adapt the directional coefficients of an existing dual-space objective. A learnable convex fusion preserves their common loss level, while a fixed absolute-gap term provides a conservative upper bound against local redistribution of the fusion coefficients. We formulate supervision allocation as a loss-based generalized-Bayesian decision problem: under specified endpoint actions, the unique control distribution and its posterior-mean action recover the implemented complementary directional weights exactly. The controller is recomputed on every mini-batch, preserves a fixed coefficient budget, and requires no additional language model. Our primary evaluation uses two LoRA-based TinyLLaMA-1.1B distillation settings with a fixed distillation temperature and dynamic weight allocation, improving average ROUGE-L over DSKD by 2.04 points for Mistral-7B $\rightarrow$ TinyLLaMA-1.1B and 2.08 points for LLaMA2-7B $\rightarrow$ TinyLLaMA-1.1B. The code will be released soon.

1 INTRODUCTION

Knowledge distillation (KD) transfers the predictive behavior of a large teacher model to a smaller student and has become a standard route toward efficient language-model deployment (Hinton et al., 2015; Kim & Rush, 2016; Gou et al., 2021; Gu et al., 2024; Xu et al., 2024). Most white-box KD objectives assume that teacher and student predictions live in compatible output spaces. This assumption breaks down for heterogeneous language-model families, where different tokenizers induce mismatched vocabulary coordinates, token boundaries, and sequence lengths. Recent cross-tokenizer distillation methods therefore focus on recovering a usable correspondence between the two predictive spaces, for example through universal logit matching, optimal-transport-based alignment, likelihood matching, or explicit dual-space transfer (Boizard et al., 2025; Cui et al., 2025; Zhang et al., 2024).

These advances primarily answer *where* knowledge can be aligned across heterogeneous output spaces. A separate question is *how much* supervision should be assigned to each aligned transfer channel during optimization. This issue is particularly relevant to dual-space distillation. DSKD (Zhang et al., 2024), for example, constructs teacher-to-student and student-to-teacher transfer paths to mitigate output-space discrepancy, but its loss terms are combined with coefficients that are fixed throughout training. Such a stationary allocation is agnostic to the current teacher–student state. More broadly, the importance of balancing heterogeneous objectives has been repeatedly observed in multi-task and multi-objective learning, where manually fixed coefficients can be sensitive to differences in loss scale and optimization dynamics, motivating adaptive or learned weighting strategies (Kendall et al., 2018; Chen et al., 2018). Related ideas have also been explored in knowledge distillation: Dynamic KD adapts the distillation procedure according to student competency (Li et al., 2021), while RW-KD learns sample-wise weights between task and distillation losses (Lu et al., 2021). These findings suggest that the useful strength of transferred supervision need not remain constant, but may vary with the sample and the evolving student state.

Figure 1 illustrates the mismatch that motivates our work. Even within a single epoch, the student encounters a wide range of fitting difficulty, while the teacher and student provide related but not

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

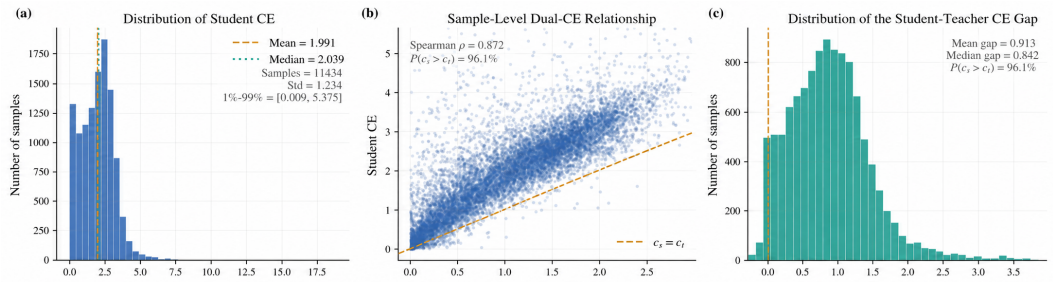


Figure 1: **Empirical motivation for state-aware supervision allocation.** Statistics are collected during the first training epoch of the Mistral-7B $\rightarrow$ TinyLLaMA-1.1B setting. (a) Student CE exhibits substantial sample-level variation rather than concentrating around a single difficulty level. (b) Student and teacher CEs are strongly correlated but non-identical ($\rho = 0.872$), and most samples satisfy $c_s > c_t$. (c) The student–teacher CE gap is broadly distributed, indicating that the relative state of the two models also varies across samples. A fixed-weight DSKD objective applies the same coefficient vector to all of these states and therefore cannot condition its supervision allocation on the observed batch difficulty or teacher–student discrepancy. The figure motivates state-conditioned allocation; it does not by itself imply that any specific adaptive rule is optimal.

redundant views of the same mini-batch. The corresponding CE discrepancy also varies substantially. Consequently, a single global coefficient vector necessarily treats heterogeneous teacher–student states identically. This limitation is orthogonal to cross-tokenizer alignment: even after the two output spaces have been bridged, the optimization procedure still requires a policy for distributing supervision across the available transfer channels. We therefore ask: *can inexpensive mini-batch evidence be used to redistribute a fixed supervision budget across dual-space distillation paths without introducing an additional controller network?*

We introduce EAKD, an evidence-adaptive dual-space distillation framework that augments the alignment backbone with an analytical weight controller. For each mini-batch, we compute detached student and teacher cross-entropies, c_s and c_t, and combine them through a learnable convex fusion together with a small absolute-gap correction,

$$d = \rho_s c_s + \rho_t c_t + \gamma |c_s - c_t|, \qquad \rho_s + \rho_t = 1, \tag{1}$$

where γ is fixed. The fused term captures the common difficulty level, whereas the gap preserves disagreement that would otherwise be attenuated by averaging. Because native-token CEs are tokenizer- and model-dependent quantities, we use them only as operational control evidence rather than as tokenizer-invariant measures of model quality.

Given this evidence, we formulate supervision allocation as a KL-regularized decision over two prescribed allocation endpoints using loss-based generalized Bayesian updating (Bissiri et al., 2016). The resulting generalized posterior has a sigmoid mixing probability, and its posterior-mean action yields complementary coefficients for the directional transfer losses under a fixed total coefficient budget. In the primary experiments, only the three dual-space loss weights are adapted per mini-batch, while the distillation temperature is fixed. The controller therefore adds no extra model forward pass, sampling procedure, or auxiliary policy network. The student model and cross-space projectors are optimized by the resulting weighted distillation objective, while gradients through the analytical coefficients update only the two scalar CE-fusion logits.

Our contributions are summarized as follows:

- **State-aware formulation of dual-space KD.** We identify a limitation left unresolved by cross-tokenizer alignment: fixed loss coefficients cannot condition supervision allocation on the evolving teacher–student state. We introduce a lightweight dual-CE evidence statistic that combines student difficulty, teacher-side evidence, and their discrepancy at the mini-batch level.
- **Analytical evidence-to-allocation controller.** We derive a bounded dynamic weighting rule from a KL-regularized generalized-Bayesian decision formulation. The posterior-mean

action redistributes a fixed coefficient budget across the T→S CE, T→S KD, and S→T KD channels, providing an interpretable alternative to an auxiliary learned controller.

- **Evaluation under a unified LoRA distillation protocol.** On Mistral-7B → TinyLLaMA-1.1B and LLaMA2-7B → TinyLLaMA-1.1B, the weight-only controller improves the DSKD average by 2.04 and 2.08 ROUGE-L points, respectively. We further report Qwen1.5-1.8B → GPT2-120M in the appendix as a more heterogeneous stress test, where the same dynamic weighting rule is paired with an additional analytical temperature controller.

2 Related Work

2.1 Cross-tokenizer and dual-space distillation.

Standard knowledge distillation typically assumes compatible teacher and student output spaces, which breaks down when the two models use different tokenizers (Hinton et al., 2015; Gu et al., 2024). Recent cross-tokenizer methods therefore focus on establishing correspondence across heterogeneous vocabularies. ULD and MultiLevelOT perform distribution matching through transport-based alignment (Boizard et al., 2025; Cui et al., 2025), while contextual mapping methods explicitly model sequence- and vocabulary-level mismatch (Chen et al., 2025). DSKD instead constructs bidirectional projections and cross-model attention to distill knowledge in both teacher and student output spaces (Zhang et al., 2024; 2025). We retain this dual-space alignment backbone and focus on a complementary problem: how to allocate supervision across the aligned objectives as the teacher–student learning state evolves. Accordingly, our method adapts the relative loss coefficients rather than redesigning the alignment mechanism. Temperature adaptation is used only in the additional Qwen1.5 → GPT2 stress test reported in the appendix.

2.2 Adaptive distillation.

Temperature, divergence, and teacher–student interpolation have previously been adapted to the learning process (Li et al., 2023; Ko et al., 2024; Shing et al., 2025). DWA-KD is a particularly relevant cross-tokenizer baseline: it combines entropy-based token weighting with Soft-DTW alignment at embedding and final hidden-state levels (Vu et al., 2026). In contrast, EAKD uses ground-truth predictive CEs as batch-level observations and adjusts coupled coefficients of existing dual-space losses. The primary method keeps the distillation temperature fixed, so its adaptation acts directly on supervision allocation; an optional temperature extension is evaluated separately in Appendix B.6. This distinction is about the statistic, granularity, and affected objective components, not the presence or absence of adaptation in prior work.

The decision-theoretic motivation follows generalized Bayesian updating, where information is connected to a distribution through a task-specific loss rather than a fully specified sampling likelihood (Bissiri et al., 2016). In parallel, robust optimization studies decision making under parameter uncertainty through explicitly specified uncertainty sets (Ben-Tal et al., 2009; Bertsimas et al., 2011), which motivates our local examination of the CE aggregation under alternative weighting choices. We use these principles only to motivate the low-dimensional control decision and local CE fusion, rather than to construct a full probabilistic or robust-optimization model. Consequently, our formulation should not be interpreted as yielding a calibrated posterior over teacher reliability or a distributionally robust training guarantee.

3 Method

Figure 2 summarizes the evidence-to-allocation loop. We first specify the inherited supervision channels, then construct gap-augmented evidence, and finally derive the analytical weight allocation from a generalized-Bayesian control decision.

3.1 Dual-Space Objective

Let $\mathcal{T}_\phi$ be the frozen teacher and $\mathcal{S}_\theta$ the trainable student. The same input–target text is tokenized independently, producing native logits Z^t, Z^s and last-layer representations H_T, H_S. Following

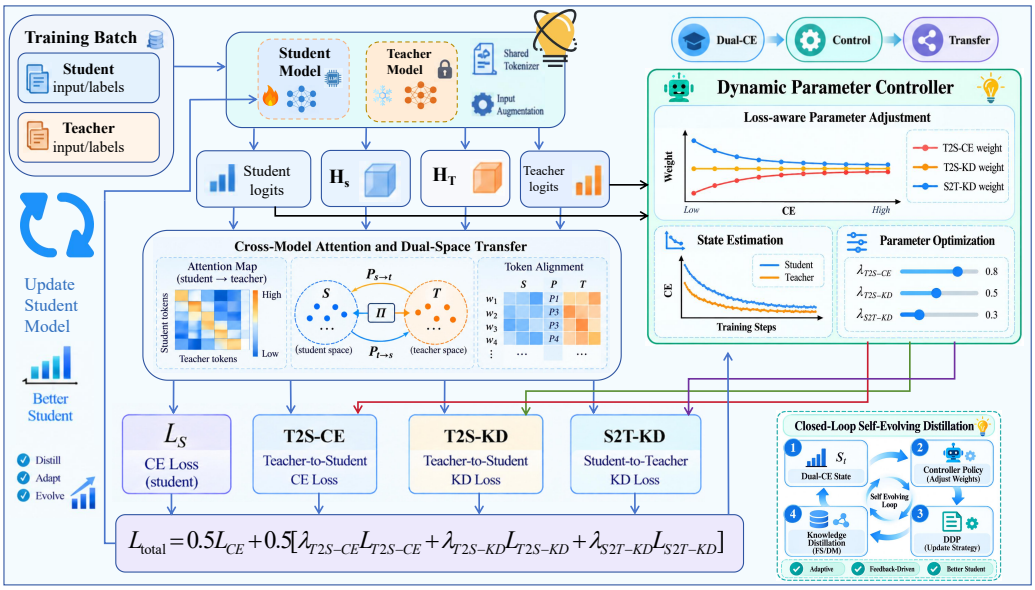


Figure 2: Overview of EAKD. A frozen teacher and a trainable student supply native predictions and hidden representations for bidirectional dual-space distillation. Detached student and teacher CEs form a gap-augmented fusion statistic. The analytical controller converts this evidence into complementary directional weights on every mini-batch while the primary distillation temperature remains fixed. The student, projectors, and fusion logits are optimized through the resulting objective.

DSKD (Zhang et al., 2024), cross-model attention aligns projected student queries and teacher keys, with padding masked before softmax. Teacher values projected into the student space yield logits $\widetilde{Z}^{t\to s}$; reverse attention and student-to-teacher projection yield $\widetilde{Z}^{s\to t}$. We denote all trainable projection parameters by ξ.

For compactness, let L_A, L_B, and L_C denote the three inherited losses:

$$
\begin{aligned}
L_A &= \mathrm{CE}_{\mathrm{sum}}(\widetilde{Z}^{t\to s}, y^s), \\
L_B &= \sum_{i,r\in\Omega_i^s} m_{ir}\, D^\star\Big(p_\tau(Z^s_{ir}), p_\tau(\mathrm{sg}[\widetilde{Z}^{t\to s}_{ir}])\Big), \\
L_C &= \sum_{i,r\in\Omega_i^t} D_{\mathrm{KL}}\Big(p_\tau(Z^t_{ir}) \,\|\, p_\tau(\widetilde{Z}^{s\to t}_{ir})\Big).
\end{aligned}
\tag{2}
$$

Here $p_\tau(z) = \mathrm{softmax}(z/\tau)$, Ω_i^m contains valid target positions, and $m_{ir} = \mathbf{1}[\arg\max \widetilde{Z}^{t\to s}_{ir} = y^s_{ir}]$ is the inherited reliable-token mask. $D^\star$ is the configured student-space discrepancy; the teacher-space discrepancy is forward KL. L_A calibrates the projected teacher against native student targets, L_B transfers a detached projected distribution to the student, and L_C constrains the projected student in the teacher space. These are valid-token sums, not three independently normalized token means.

The supervised student loss is $L_{\mathrm{CE}} = \mathrm{CE}_{\mathrm{sum}}(Z^s, y^s)$. Let $N_{\mathcal{B}} > 0$ be the shared batch normalizer supplied by the training driver. The objective is

$$
J_{\mathcal{B}} = \frac{(1-\lambda_{\mathrm{KD}})L_{\mathrm{CE}} + \lambda_{\mathrm{KD}}(w_A L_A + w_B L_B + w_C L_C)}{N_{\mathcal{B}}}.
\tag{3}
$$

The controller changes only the three coefficients without changing these inherited loss definitions. In the primary experiments, the shared distillation temperature is fixed to $\tau_0 = 2.0$; Appendix B.6 describes the temperature-augmented Qwen1.5 $\to$ GPT2 extension.

3.2 DUAL-CE EVIDENCE AND ROBUST FUSION

For model $m \in \{s, t\}$, let Ω_i^m contain the valid target positions of example i. We average its mean target-token CE over the batch and detach the observation:

$$\widehat{c}_m = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{|\Omega_i^m|}\sum_{r\in\Omega_i^m} -\log p_m(y_{ir}^m \mid x_i^m, y_{i,<r}^m), \qquad c_m = \mathrm{sg}[\widehat{c}_m]. \tag{4}$$

Under data parallelism, local means are averaged across equal-sized per-rank batches before control is computed. Thus the gap below is the absolute difference of synchronized batch means, not a mean of per-example absolute gaps.

Two trainable logits define the nominal fusion coefficients, and a fixed gap correction gives the control statistic:

$$\begin{aligned}[\rho_s, \rho_t] = \mathrm{softmax}([\ell_s, \ell_t]), \qquad d_0 = \rho_s c_s + \rho_t c_t,\\ g = |c_s - c_t|, \qquad d = d_0 + \gamma g, \qquad \gamma = 0.1.\end{aligned} \tag{5}$$

The logits belong to the optimized distillation model; only their difference determines the fusion. The gap modifies the control evidence and is not a separate penalty in Eq. (3).

To interpret the correction, consider redistributing fusion mass by δ while keeping the coefficients non-negative and normalized:

$$\mathcal{U}_\gamma(\rho) = \{(\rho_s + \delta, \rho_t - \delta) : |\delta| \le \gamma,\ -\rho_s \le \delta \le \rho_t\}. \tag{6}$$

This is a total-variation neighborhood of radius γ over the two fusion sources, not an uncertainty set over all training examples.

Proposition 1 (Gap-augmented fusion bound)**.** *For $\rho_s, \rho_t \ge 0$, $\rho_s + \rho_t = 1$, and $\gamma \ge 0$, define*

$$R_\gamma(c; \rho) = \max_{v\in\mathcal{U}_\gamma(\rho)} (v_s c_s + v_t c_t). \tag{7}$$

Then $d_0 \le R_\gamma(c;\rho) \le d_0 + \gamma|c_s - c_t| = d$. In particular, $R_\gamma = d$ whenever $\min(\rho_s, \rho_t) \ge \gamma$.

The bound follows by maximizing $d_0 + \delta(c_s - c_t)$; Appendix A.2 gives the exact boundary cases. When the observed CEs agree, redistributing fusion mass has no effect. When they differ, the correction provides a conservative upper envelope of the resulting change. Softmax does not guarantee $\rho_s, \rho_t \ge \gamma$, so we use the general bound rather than assume equality. The radius 0.1 is a modeling choice, not a statistically calibrated confidence radius. Native-token CE scales may differ across tokenizers; the statistic describes the observed training scale, not teacher superiority or a predictive-distribution divergence.

3.3 GENERALIZED-BAYESIAN ANALYTICAL WEIGHT CONTROL

We now connect the evidence to supervision allocation. Let $Z \in \{0, 1\}$ index two prescribed allocation endpoints and let u be the evidence for that decision. We use a loss-based update (Bissiri et al., 2016):

$$q_u^\star = \arg\min_{q\in\Delta_2}\left\{\mathbb{E}_{Z\sim q}[\mathcal{E}(Z;u)] + \frac{1}{\beta}D_{\mathrm{KL}}(q\|\pi)\right\}, \quad \mathcal{E}(0;u) = u, \quad \mathcal{E}(1;u) = 0, \tag{8}$$

where Δ_2 is the probability simplex over two endpoints, $\pi \in \Delta_2$ has strictly positive entries, and $\beta > 0$ sets the evidence scale. The energy difference encodes the chosen direction: larger evidence favors endpoint 1. It is a structural control assumption, not a likelihood model for CE observations.

Proposition 2 (Analytical generalized-Bayes action)**.** *For finite u, the unique minimizer of Eq.* (8) *satisfies*

$$q_u^\star(z) \propto \pi_z e^{-\beta\mathcal{E}(z;u)}, \qquad r_u = q_u^\star(1) = \sigma(b + \beta u), \qquad b = \log(\pi_1/\pi_0). \tag{9}$$

For endpoint actions $\mathbf{a}_0, \mathbf{a}_1$, the unique minimizer of $\mathbb{E}_{q_u^\star}\|\mathbf{a} - \mathbf{a}_Z\|_2^2$ is $\mathbf{a}^\star = (1 - r_u)\mathbf{a}_0 + r_u\mathbf{a}_1$.

Appendix A.1 proves the result and the associated control-objective optimality identity. These are posterior-mean actions, not MAP estimates of Z.

For the primary controller we set $u = d$, $\pi = (1/2, 1/2)$, and $\beta = 1$. For weights ordered as (w_A, w_B, w_C), the two endpoint actions are

$$\mathbf{a}_0 = (w_{\max}, w_{\mathrm{mid}}, w_{\min}), \qquad \mathbf{a}_1 = (w_{\min}, w_{\mathrm{mid}}, w_{\max}), \tag{10}$$

where $w_{\mathrm{mid}} = (w_{\min} + w_{\max})/2$. Writing $\Delta_w = w_{\max} - w_{\min}$, Proposition 2 yields

$$r = \sigma(d), \qquad w_A = w_{\max} - \Delta_w r, \quad w_B = w_{\mathrm{mid}}, \quad w_C = w_{\min} + \Delta_w r. \tag{11}$$

We use $(w_{\min}, w_{\max}) = (0.1, 10)$, so $w_B = 5.05$. The endpoint ordering encodes our hypothesis that larger fused evidence should shift relative weight from projected teacher calibration toward teacher-space matching. The empirical comparisons test this design; the decision theorem does not prove it universally preferable. Neither Z nor r is a teacher-correctness indicator.

The gap has an exact interpretation in the control distribution. If $r_0 = \sigma(d_0)$ and $r_\gamma = \sigma(d)$, then

$$\mathrm{logit}(r_\gamma) - \mathrm{logit}(r_0) = \gamma |c_s - c_t|. \tag{12}$$

Thus the robust upper envelope changes the generalized log-odds by a specified amount rather than adding an unrelated student-loss penalty. The posterior-mean construction also preserves $w_A + w_C = w_{\max} + w_{\min}$ and total coefficient $3w_{\mathrm{mid}}$; its sensitivity and reachable ranges are derived in Appendix A.

The decision is recomputed from the same fixed prior on every mini-batch, not accumulated as a posterior across training. No sampling or inner numerical optimizer is required. The CE observations are detached, while the weight formulas remain differentiable in (ℓ_s, ℓ_t). In particular,

$$\begin{aligned} \frac{\partial J_{\mathcal{B}}}{\partial \ell_s} &= \frac{\lambda_{\mathrm{KD}} \Delta_w}{N_{\mathcal{B}}} r(1-r)(L_C - L_A)\rho_s \rho_t (c_s - c_t), \\ \frac{\partial J_{\mathcal{B}}}{\partial \ell_t} &= -\frac{\partial J_{\mathcal{B}}}{\partial \ell_s}. \end{aligned} \tag{13}$$

The fusion is therefore trained through the actual weighted KD objective, not by empirical-Bayes marginal likelihood or a surrogate text-generation gradient. Appendix A.3 provides the derivation. This yields one coherent loop: dual-CE evidence with the gap correction determines a single allocation distribution, its mean action specifies the directional weights, and the student and fusion updates change the next batch's evidence. The optional temperature decision used only for Qwen1.5 $\rightarrow$ GPT2 is described in Appendix B.6.

4 Experiments

The analysis establishes the form and properties of the controller under explicit assumptions. The experiments assess whether this allocation is useful for the chosen dual-space backbone; conditional decision optimality is not treated as a prediction of test performance.

4.1 Experimental Setup

Our primary evaluation uses two LoRA-based TinyLLaMA-1.1B distillation settings: Mistral-7B $\rightarrow$ TinyLLaMA-1.1B and the shared-vocabulary LLaMA2-7B $\rightarrow$ TinyLLaMA-1.1B configuration. Training uses the preprocessed Databricks-Dolly-15k (Conover et al., 2023) split with more than 11,000 training examples, 1,000 validation examples, and 500 test examples. Evaluation covers Dolly (500 examples) (Conover et al., 2023), SelfInst (242) (Wang et al., 2023), VicunaEval (80) (Chiang et al., 2023), S-NI (1,694) (Wang et al., 2022), and UnNI (10,000) (Honovich et al., 2023). We report ROUGE-L (Lin, 2004); the Average column weights the five evaluation sets equally. Results are reported as means and standard deviations over five runs with seeds $\{10, 20, 30, 40, 50\}$. Decoding temperature and top-p are both 1.0.

Experiments use three NVIDIA A100 80GB GPUs. TinyLLaMA, LLaMA2, and Mistral checkpoints are adapted for 10 epochs at learning rate 10^{-3} with batch sizes 4, 8, and 8, respectively. Projectors use learning rate 10^{-3} with cosine decay. TinyLLaMA, LLaMA2, and Mistral use LoRA with rank 256, alpha 8, and dropout 0.1. Teachers remain frozen during distillation.

The primary EAKD protocol uses the mini-batch evidence in Eq. (5) and the analytical weights in Eq. (11), with $\gamma = 0.1$, coefficient bounds $[0.1, 10]$, and a fixed shared distillation temperature

$\tau_0 = 2.0$. The generalized allocation distribution uses a uniform prior and unit evidence scale. CE observations are synchronized across data-parallel ranks before the control variable is computed. The fusion logits are optimized with the trainable distillation components, and the weight action is recomputed for every mini-batch. Appendix C provides the implementation procedure.

4.2 Main results

Table 1 compares EAKD with sequence-level, distribution-level, and dual-space distillation baselines under the two primary TinyLLaMA-1.1B settings. For Mistral-7B → TinyLLaMA-1.1B, EAKD improves the DSKD average from 24.86 to 26.90, corresponding to a gain of 2.04 ROUGE-L points. A similar improvement is observed in the shared-vocabulary LLaMA2-7B → TinyLLaMA-1.1B setting, where the average increases from 25.05 to 27.13 (+2.08). Importantly, these gains are not driven by a single evaluation set: EAKD improves over DSKD on all five datasets in both primary teacher–student configurations. The comparable gains across the different-vocabulary Mistral setting and the shared-vocabulary LLaMA2 setting suggest that adaptive supervision allocation remains useful under different degrees of teacher–student output-space compatibility.

Table 1: ROUGE-L (↑) across five evaluation sets for the two primary LoRA-based TinyLLaMA distillation settings. Results are means with standard deviations in subscript. Green values in the Average column indicate the absolute gain of EAKD over DSKD.

Methods	Dolly	SelfInst	VicunaEval	S-NI	UnNI	Average
Mistral-7B → TinyLLaMA-1.1B						
Teacher	$31.56_{\pm 0.19}$	$25.10_{\pm 0.36}$	$20.50_{\pm 0.32}$	$36.07_{\pm 0.24}$	$36.27_{\pm 0.15}$	29.90
SeqKD (Kim & Rush, 2016)	$23.56_{\pm 0.39}$	$15.87_{\pm 0.54}$	$15.99_{\pm 0.55}$	$25.50_{\pm 0.37}$	$26.64_{\pm 0.09}$	21.51
MinED (Wan et al., 2024)	$20.96_{\pm 0.51}$	$14.49_{\pm 0.35}$	$15.98_{\pm 0.45}$	$27.21_{\pm 0.13}$	$26.47_{\pm 0.11}$	21.77
ULD (Boizard et al., 2025)	$22.80_{\pm 0.28}$	$15.93_{\pm 0.74}$	$16.43_{\pm 0.60}$	$26.94_{\pm 0.28}$	$24.83_{\pm 0.13}$	20.64
MultiLevelOT (Cui et al., 2025)	$23.84_{\pm 0.58}$	$17.64_{\pm 0.59}$	$15.74_{\pm 0.35}$	$29.04_{\pm 0.35}$	$29.26_{\pm 0.16}$	23.10
DSKDv2 (Zhang et al., 2025)	$26.17_{\pm 0.25}$	$18.86_{\pm 0.89}$	$18.17_{\pm 0.27}$	$31.54_{\pm 0.30}$	$30.57_{\pm 0.16}$	25.06
DSKD (Zhang et al., 2024)	$26.52_{\pm 0.45}$	$17.90_{\pm 0.69}$	$18.20_{\pm 0.59}$	$30.66_{\pm 0.39}$	$31.03_{\pm 0.11}$	24.86
DWA-KD (Vu et al., 2026)	$26.67_{\pm 0.36}$	$18.48_{\pm 0.67}$	$18.93_{\pm 0.31}$	$30.32_{\pm 0.21}$	30.92 ± 0.21	25.06
EAKD	$\mathbf{27.73}_{\pm 0.25}$	$\mathbf{19.61}_{\pm 0.51}$	$\mathbf{20.63}_{\pm 0.42}$	$\mathbf{33.45}_{\pm 0.17}$	$\mathbf{33.07}_{\pm 0.08}$	**26.90** (+2.04↑)
LLaMA2-7B → TinyLLaMA-1.1B (Same Vocabulary)						
Teacher	$28.32_{\pm 0.46}$	$20.95_{\pm 0.69}$	$18.76_{\pm 0.35}$	$32.05_{\pm 0.28}$	$32.41_{\pm 0.12}$	26.50
SFT	$23.20_{\pm 0.13}$	$14.88_{\pm 0.54}$	$16.42_{\pm 0.35}$	$27.79_{\pm 0.27}$	$26.12_{\pm 0.11}$	21.68
SeqKD (Kim & Rush, 2016)	$23.21_{\pm 0.22}$	$16.46_{\pm 0.72}$	$16.58_{\pm 0.38}$	$26.33_{\pm 0.26}$	$27.69_{\pm 0.10}$	22.05
GKD (Agarwal et al., 2024)	$23.45_{\pm 0.60}$	$16.70_{\pm 0.30}$	$15.61_{\pm 0.51}$	$28.18_{\pm 0.18}$	$27.60_{\pm 0.23}$	22.31
MiniLLM (Gu et al., 2024)	$23.28_{\pm 0.49}$	$17.70_{\pm 0.46}$	$15.34_{\pm 0.52}$	$31.57_{\pm 0.29}$	$29.40_{\pm 0.08}$	23.46
DistiLLM (Ko et al., 2024)	$24.63_{\pm 0.32}$	$17.43_{\pm 0.83}$	$16.29_{\pm 0.25}$	$31.07_{\pm 0.23}$	$29.85_{\pm 0.17}$	23.85
DSKDv2 (Zhang et al., 2025)	$26.80_{\pm 0.57}$	$18.81_{\pm 1.17}$	$17.86_{\pm 0.11}$	$31.90_{\pm 0.17}$	$31.80_{\pm 0.08}$	25.43
DSKD (Zhang et al., 2024)	$26.31_{\pm 0.26}$	$18.27_{\pm 0.56}$	$18.04_{\pm 0.37}$	$31.43_{\pm 0.26}$	$31.20_{\pm 0.09}$	25.05
DWA-KD (Vu et al., 2026)	$26.46_{\pm 0.25}$	$18.57_{\pm 0.52}$	$19.11_{\pm 0.27}$	$32.21_{\pm 0.16}$	31.44 ± 0.17	25.55
EAKD	$\mathbf{27.93}_{\pm 0.45}$	$\mathbf{20.92}_{\pm 0.32}$	$\mathbf{20.22}_{\pm 0.32}$	$\mathbf{33.51}_{\pm 0.21}$	$\mathbf{33.07}_{\pm 0.11}$	**27.13** (+2.08↑)

4.3 Further Analysis

Robustness to discrepancy functions. Table 2 further examines whether the benefit of EAKD depends on the particular discrepancy used for student-space distribution matching. We consider forward KL (Kullback & Leibler, 1951), reverse KL (RKL) (Gu et al., 2024), Jensen–Shannon divergence (JS) (Lin, 1991), and the skew variants SKL and SRKL (Ko et al., 2024). For both DSKD and EAKD, the selected discrepancy is applied in the student space while forward KL is retained in the teacher space, so that only the student-space matching objective is varied.

For Mistral-7B → TinyLLaMA-1.1B, EAKD improves the corresponding DSKD average under all five discrepancies, with gains ranging from 0.53 to 2.63 ROUGE-L points. The same aggregate trend is observed for LLaMA2-7B → TinyLLaMA-1.1B, where EAKD improves over DSKD by 2.08, 1.45, 2.14, 2.76, and 2.33 points for KL, RKL, JS, SKL, and SRKL, respectively. The gain magnitude varies with the discrepancy, and individual dataset–loss combinations do not always improve, indicating that the underlying distribution-matching objective still affects the final behavior. Nevertheless, the average improves for every tested discrepancy in both teacher–student settings. Since the evidence-to-allocation controller is kept unchanged across these experiments, the results

suggest that EAKD is not tied to a particular student-space divergence and can be applied on top of different KL-family distillation objectives.

Table 2: Comparison of student-space discrepancy functions across the two primary TinyLLaMA-1.1B distillation settings. For DSKD and EAKD, the specified discrepancy is applied in the student space while forward KL is retained in the teacher space. Parenthesized values denote absolute improvements of EAKD over the corresponding DSKD average.

Loss	Method	Dolly	SelfInst	VicunaEval	S-NI	UnNI	Avg.
Mistral-7B → TinyLLaMA-1.1B							
KL	DSKD	$26.52_{\pm 0.45}$	$17.90_{\pm 0.69}$	$18.20_{\pm 0.59}$	$30.66_{\pm 0.39}$	$31.03_{\pm 0.11}$	24.86
	EAKD	$\mathbf{27.73}_{\pm 0.25}$	$\mathbf{19.61}_{\pm 0.51}$	$\mathbf{20.63}_{\pm 0.42}$	$\mathbf{33.45}_{\pm 0.17}$	$\mathbf{33.07}_{\pm 0.08}$	**26.90** (+2.04↑)
RKL	DSKD	$25.41_{\pm 0.18}$	$18.31_{\pm 0.45}$	$16.83_{\pm 0.46}$	$\mathbf{34.79}_{\pm 0.16}$	$\mathbf{34.05}_{\pm 0.12}$	25.88
	EAKD	$\mathbf{27.68}_{\pm 0.68}$	$\mathbf{20.11}_{\pm 0.47}$	$\mathbf{20.82}_{\pm 0.41}$	$32.93_{\pm 0.14}$	$33.52_{\pm 0.14}$	**27.01** (+1.13↑)
JS	DSKD	$24.09_{\pm 0.71}$	$16.77_{\pm 0.75}$	$16.96_{\pm 0.27}$	$30.01_{\pm 0.15}$	$30.00_{\pm 0.10}$	23.56
	EAKD	$\mathbf{27.63}_{\pm 0.05}$	$\mathbf{18.68}_{\pm 0.26}$	$\mathbf{19.65}_{\pm 0.36}$	$\mathbf{32.23}_{\pm 0.08}$	$\mathbf{32.78}_{\pm 0.01}$	**26.19** (+2.63↑)
SKL	DSKD	$\mathbf{25.28}_{\pm 0.24}$	$\mathbf{17.33}_{\pm 0.62}$	$17.57_{\pm 0.43}$	$30.27_{\pm 0.30}$	$31.14_{\pm 0.35}$	24.32
	EAKD	$24.47_{\pm 0.24}$	$15.58_{\pm 0.43}$	$\mathbf{18.46}_{\pm 0.07}$	$\mathbf{32.77}_{\pm 0.05}$	$\mathbf{32.95}_{\pm 0.01}$	**24.85** (+0.53↑)
SRKL	DSKD	$24.87_{\pm 0.50}$	$17.63_{\pm 0.53}$	$17.16_{\pm 0.24}$	$29.77_{\pm 0.19}$	$30.78_{\pm 0.14}$	24.04
	EAKD	$\mathbf{27.50}_{\pm 0.37}$	$\mathbf{20.39}_{\pm 0.52}$	$\mathbf{18.58}_{\pm 0.06}$	$\mathbf{32.87}_{\pm 0.29}$	$\mathbf{32.91}_{\pm 0.05}$	**26.45** (+2.41↑)
LLaMA2-7B → TinyLLaMA-1.1B (Same Vocabulary)							
KL	Original	$25.46_{\pm 0.63}$	$17.21_{\pm 0.25}$	$16.43_{\pm 0.53}$	$29.27_{\pm 0.29}$	$29.28_{\pm 0.09}$	23.53
	DSKD	$26.31_{\pm 0.26}$	$18.27_{\pm 0.56}$	$18.04_{\pm 0.37}$	$31.43_{\pm 0.26}$	$31.20_{\pm 0.09}$	25.05
	EAKD	$\mathbf{27.93}_{\pm 0.45}$	$\mathbf{20.92}_{\pm 0.32}$	$\mathbf{20.22}_{\pm 0.32}$	$\mathbf{33.51}_{\pm 0.21}$	$\mathbf{33.07}_{\pm 0.11}$	**27.13** (+2.08↑)
RKL	Original	$24.49_{\pm 0.41}$	$17.14_{\pm 0.61}$	$16.87_{\pm 0.26}$	$29.50_{\pm 0.28}$	$29.36_{\pm 0.08}$	23.47
	DSKD	$\mathbf{26.93}_{\pm 0.34}$	$18.14_{\pm 0.54}$	$\mathbf{18.81}_{\pm 0.39}$	$31.79_{\pm 0.31}$	$32.49_{\pm 0.11}$	25.63
	EAKD	$26.67_{\pm 0.15}$	$\mathbf{19.21}_{\pm 0.51}$	$17.62_{\pm 0.24}$	$\mathbf{36.09}_{\pm 0.05}$	$\mathbf{35.82}_{\pm 0.07}$	**27.08** (+1.45↑)
JS	Original	$24.03_{\pm 0.31}$	$15.75_{\pm 0.51}$	$16.64_{\pm 0.30}$	$28.08_{\pm 0.10}$	$28.68_{\pm 0.08}$	22.62
	DSKD	$24.79_{\pm 0.42}$	$17.10_{\pm 0.47}$	$\mathbf{16.78}_{\pm 0.20}$	$29.06_{\pm 0.18}$	$29.47_{\pm 0.22}$	23.44
	EAKD	$\mathbf{25.16}_{\pm 0.37}$	$\mathbf{19.15}_{\pm 0.63}$	$16.26_{\pm 0.41}$	$\mathbf{34.57}_{\pm 0.31}$	$\mathbf{32.78}_{\pm 0.14}$	**25.58** (+2.14↑)
SKL	Original	$24.14_{\pm 0.53}$	$15.98_{\pm 0.72}$	$16.89_{\pm 0.22}$	$29.30_{\pm 0.18}$	$28.71_{\pm 0.12}$	23.01
	DSKD	$25.88_{\pm 0.22}$	$17.59_{\pm 0.56}$	$17.17_{\pm 0.34}$	$29.52_{\pm 0.33}$	$30.69_{\pm 0.16}$	24.17
	EAKD	$\mathbf{26.12}_{\pm 0.31}$	$\mathbf{20.51}_{\pm 0.48}$	$\mathbf{18.33}_{\pm 0.27}$	$\mathbf{35.02}_{\pm 0.16}$	$\mathbf{34.68}_{\pm 0.09}$	**26.93** (+2.76↑)
SRKL	Original	$24.28_{\pm 0.58}$	$16.91_{\pm 0.67}$	$16.88_{\pm 0.20}$	$29.55_{\pm 0.19}$	$28.64_{\pm 0.21}$	23.25
	DSKD	$25.44_{\pm 0.22}$	$17.34_{\pm 0.69}$	$17.19_{\pm 0.34}$	$30.29_{\pm 0.29}$	$31.23_{\pm 0.13}$	24.30
	EAKD	$\mathbf{26.29}_{\pm 0.41}$	$\mathbf{20.03}_{\pm 0.08}$	$\mathbf{18.39}_{\pm 0.49}$	$\mathbf{34.82}_{\pm 0.32}$	$\mathbf{33.64}_{\pm 0.06}$	**26.63** (+2.33↑)

Representation and response diagnostics.

Representation structure analysis. We further examine whether the improvement in generation quality is accompanied by a closer teacher–student representation structure. Figure 3 reports the sample-level structure distances for Mistral-7B → TinyLLaMA-1.1B under two relational structure measures: cosine similarity and normalized inner product. Because the teacher and student hidden dimensions differ, we compare token–token relational structures rather than hidden vectors directly.

As shown in Fig. 3(a), EAKD produces a clear downward shift in the cosine-structure distance distribution relative to both SFT and DSKD, with the median decreasing from approximately 0.50 for SFT and 0.49 for DSKD to approximately 0.45 for EAKD. A similar trend is observed for the normalized inner-product structure in Fig. 3(b), where EAKD reduces the median distance to approximately 0.49, compared with approximately 0.54–0.55 for SFT and DSKD. The lower medians and downward-shifted distributions under both structure definitions indicate that the student trained with adaptive supervision allocation preserves a relational geometry that is closer to the teacher.

The precise definitions of the cosine and normalized-inner-product structure matrices, together with the corresponding distance functions, are provided in Appendix B.3. We additionally report the same analysis for LLaMA2-7B → TinyLLaMA-1.1B in the appendix, providing a complementary shared-tokenization setting.

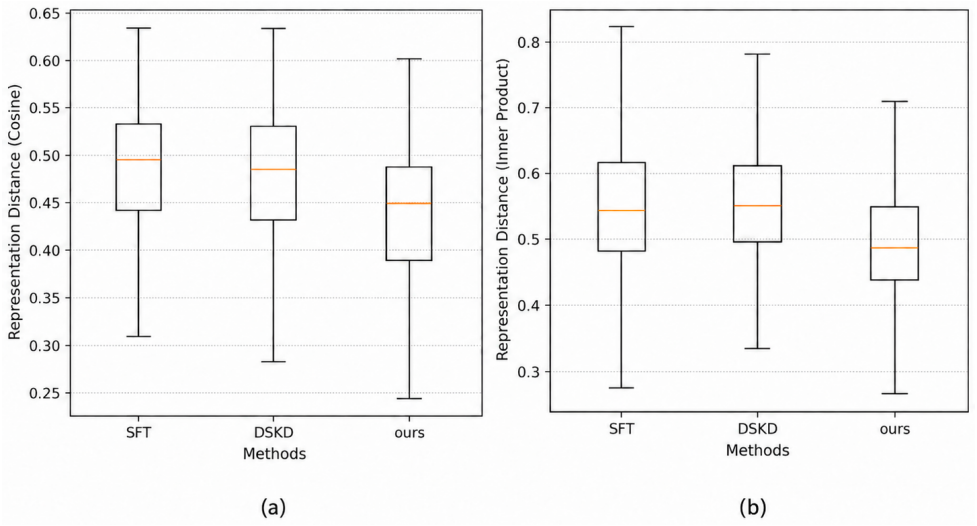


Figure 3: Teacher–student representation-structure distance distributions for Mistral-7B → TinyLLaMA-1.1B on Dolly. We compare SFT, DSKD, and EAKD using (a) cosine-based relational structure and (b) normalized-inner-product relational structure. Each box summarizes sample-level distances, where a lower value indicates a closer match between the student and teacher relational structures. EAKD exhibits lower median distances and an overall downward-shifted distribution under both measures.

5 Conclusion

We presented EAKD, an evidence-adaptive weight controller for dual-space language-model distillation. Learned dual-CE fusion and an absolute-gap correction produce a mini-batch evidence statistic with a local reweighting upper-bound interpretation. A specified generalized-Bayesian decision then yields complementary directional coefficients as a posterior-mean action while preserving a fixed coefficient budget. The primary protocol keeps the distillation temperature fixed and adapts only the supervision weights. Under the two unified LoRA-based TinyLLaMA settings, EAKD improves average ROUGE-L over DSKD by 2.04 and 2.08 points. The appendix further evaluates a more heterogeneous Qwen1.5 → GPT2 setting that additionally activates analytical temperature adaptation. The theory explains the analytical mechanism and its scope, while the experiments assess its utility across transfer channels, discrepancies, parameter ranges, and model-pair regimes.

Reproducibility Statement

Equations (4)–(11) specify the executed primary weight controller; Appendix B.6 specifies the additional temperature action used only for Qwen1.5 → GPT2. Appendix A supplies proofs, boundary conditions, and exact gradient derivations; Appendix C gives the update procedure and loss conventions. Experimental settings and evaluation sizes are reported in Section 4.1.

AI Use Statement

An AI writing assistant was used to help reorganize the manuscript, edit language and LaTeX, and formulate and check analytical derivations. Empirical measurements were supplied by the authors rather than generated by the assistant. The authors are responsible for checking equations, assumptions, references, implementation correspondence, and final claims.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-

generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, 2009. doi: 10.1515/9781400831050.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011. doi: 10.1137/080734510.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016. doi: 10.1111/rssb.12158.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: The universal logit distillation loss for llms. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=buBxXiG09A.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Enhancing cross-tokenizer knowledge distillation with contextual dynamical mapping. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8005–8018. Association for Computational Linguistics, 2025.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/. LMSYS Org.

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 instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm. Databricks.

Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23724–23732, 2025. doi: 10.1609/aaai.v39i22.34543.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. doi: 10.1007/s11263-021-01453-z.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL https://arxiv.org/abs/1503.02531.

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)*, pp. 14409–14428, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.806. URL https://aclanthology.org/2023.acl-long.806/.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018. doi: 10.1109/CVPR.2018.00781.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 24872–24895. PMLR, 2024.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. Dynamic knowledge distillation for pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.31.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1504–1512, 2023. doi: 10.1609/aaai.v37i2.25236.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Peng Lu, Abbas Ghaddar, Ahmad Rashid, Mehdi Rezagholizadeh, Ali Ghodsi, and Philippe Langlais. RW-KD: Sample-wise loss terms re-weighting for knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3145–3152. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.270.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. TAID: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Duc Trung Vu, Pham Khanh Chi, Dat Phi Van, Linh Ngo Van, Dinh Viet Sang, and Trung Le. DWA-KD: Dual-space weighting and time-warped alignment for cross-tokenizer knowledge distillation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3513–3527, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.181. URL https://aclanthology.org/2026.findings-eacl.181/.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge fusion of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jiDsk12qcz.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL https://aclanthology.org/2022.emnlp-main.340/.

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)*, pp. 13484–13508, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

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. *arXiv preprint arXiv:2402.13116*, 2024.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1010.

Xue Zhang, Songming Zhang, Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. A dual-space framework for general knowledge distillation of large language models. *arXiv preprint arXiv:2504.11426*, 2025.

A Theoretical Analysis and Proofs

The analysis distinguishes three objects: the observed CE pair, a decision distribution over prescribed control endpoints, and the neural training objective. The generalized posterior belongs to the second object. It is not a posterior over teacher correctness or student parameters, and the fusion uncertainty set is not a confidence set for an unknown data-generating distribution.

A.1 Generalized-Bayesian Control and Posterior-Mean Actions

Fix finite u, a strictly positive prior π, and $\beta > 0$. Writing $r = q(Z = 1)$ turns Eq. (8) into

$$F(r; u) = (1 - r)u + \frac{1}{\beta}\left[r \log \frac{r}{\pi_1} + (1 - r) \log \frac{1 - r}{\pi_0}\right], \qquad 0 \le r \le 1, \tag{14}$$

with the usual convention $0 \log 0 = 0$. On the interior,

$$\frac{\partial F}{\partial r} = -u + \frac{1}{\beta}\left[\log \frac{r}{1 - r} - \log \frac{\pi_1}{\pi_0}\right],$$
$$\frac{\partial^2 F}{\partial r^2} = \frac{1}{\beta r(1 - r)} > 0. \tag{15}$$

The first derivative tends to $-\infty$ and $+\infty$ at the two boundaries, so the unique global minimum is interior. Solving its first-order condition yields $r_u = \sigma(b + \beta u)$, where $b = \log(\pi_1/\pi_0)$. Equivalently,

$$q_u^\star(z) = \frac{\pi_z e^{-\beta \mathcal{E}(z;u)}}{Z_u}, \qquad Z_u = \pi_0 e^{-\beta u} + \pi_1. \tag{16}$$

This is the loss-based Gibbs update (Bissiri et al., 2016), applied here to a prescribed control energy. For any distribution q on the two endpoints, substitution gives the exact optimality identity

$$\mathbb{E}_q[\mathcal{E}(Z; u)] + \frac{1}{\beta} D_{\mathrm{KL}}(q \| \pi) = -\frac{1}{\beta} \log Z_u + \frac{1}{\beta} D_{\mathrm{KL}}(q \| q_u^\star). \tag{17}$$

Consequently, $F(r; u) - F(r_u; u) = \beta^{-1} D_{\mathrm{KL}}(\mathrm{Bern}(r) \| \mathrm{Bern}(r_u)) \ge 0$. Also $F'' \ge 4/\beta$, giving the conditional bound $F(r; u) - F(r_u; u) \ge 2(r - r_u)^2/\beta$. These statements concern the specified control objective, not the neural training risk.

For actions in a Euclidean space, let $\boldsymbol{\mu} = \mathbb{E}_{q_u^\star}[\mathbf{a}_Z]$. The identity

$$\mathbb{E}_{q_u^\star} \|\mathbf{a} - \mathbf{a}_Z\|_2^2 = \|\mathbf{a} - \boldsymbol{\mu}\|_2^2 + \mathbb{E}_{q_u^\star} \|\mathbf{a}_Z - \boldsymbol{\mu}\|_2^2 \tag{18}$$

proves that the unique squared-loss Bayes action is $\boldsymbol{\mu} = (1 - r_u)\mathbf{a}_0 + r_u \mathbf{a}_1$. The mean lies on the segment joining the two endpoints, so constraining the action to this segment gives the same solution. Substituting Eq. (10) proves Eq. (11). The same posterior-mean construction, applied to scalar temperature endpoints, gives the optional Qwen1.5 $\to$ GPT2 temperature action in Eq. (47). By contrast, a MAP decision about Z selects an endpoint and does not reproduce either interpolation.

The primary implementation fixes $\pi = (1/2, 1/2)$ and $\beta = 1$ for the directional-weight decision, giving $r = \sigma(d)$. The Qwen1.5 $\to$ GPT2 appendix extension uses a second decision with $u = c_s$ for temperature. The scale β is distinct from the distillation temperature τ, the KD mixing coefficient λ_{KD}, and the gap radius γ. Only the product of β and an arbitrary rescaling of the chosen energy enters the posterior odds; setting $\beta = 1$ matches the implemented CE scale but is not a theorem about the best empirical scale. Priors are reset to the same specified values at each batch. There is no sequential multiplication of batch likelihoods or accumulation of posterior odds, and no random regime is sampled.

The ordering of endpoints and the energy difference $\mathcal{E}(1; u) - \mathcal{E}(0; u) = -u$ encode a structural hypothesis about supervision allocation. Without that hypothesis, the Bayesian construction does not decide which channel should receive greater weight. The generalized control distribution is operational, not a fitted generative model for the CE statistics. Learned fusion is optimized by the KD objective and is not empirical-Bayes estimation.

Because $c_s, c_t \ge 0$, $\rho_s, \rho_t \ge 0$, and $\gamma \ge 0$, the fused evidence d is non-negative. Under the implemented uniform prior, $1/2 \le r < 1$ for finite evidence. Hence

$$w_A \in (w_{\min}, w_{\mathrm{mid}}], \qquad w_B = w_{\mathrm{mid}}, \qquad w_C \in [w_{\mathrm{mid}}, w_{\max}). \tag{19}$$

Floating-point saturation can reach a limiting endpoint. Accordingly, Z indexes allocation endpoints, not an easy–hard classifier spanning a symmetric probability range. For the optional Qwen1.5 $\rightarrow$ GPT2 temperature extension, non-negative student CE similarly restricts the nominal $[1.8, 2.2]$ interval to the reachable half-interval $[2.0, 2.2)$. Adding the coefficients gives

$$w_A + w_C = w_{\max} + w_{\min}, \qquad w_A + w_B + w_C = 3w_{\mathrm{mid}}. \tag{20}$$

The implemented total is 15.15, while three unit coefficients sum to 3. This fixed internal budget neither matches unit-weight DSKD's scale nor fixes weighted loss magnitudes or gradient norms. A scale-matched static comparison would use all three coefficients equal to 5.05.

The following identity concerns only the temperature-augmented Qwen1.5 $\rightarrow$ GPT2 experiment. For fixed logits z and $p_i(\tau) \propto \exp(z_i/\tau)$, let $Z(\tau) = \sum_i \exp(z_i/\tau)$. Since $H(p_\tau) = \log Z(\tau) - \mathbb{E}_{p_\tau}[z]/\tau$, $d \log Z/d\tau = -\mathbb{E}[z]/\tau^2$, and $d\mathbb{E}[z]/d\tau = -\operatorname{Var}(z)/\tau^2$, direct differentiation gives

$$\frac{dH(p_\tau)}{d\tau} = \frac{\operatorname{Var}_{p_\tau}(z)}{\tau^3} \geq 0. \tag{21}$$

The identity explains distributional softening at fixed logits; it does not prove that increasing temperature always improves distillation.

A.2 Gap-Augmented Fusion and Evidence Sensitivity

Let $h = c_s - c_t$ and $d_0 = \rho_s c_s + \rho_t c_t$. A feasible redistribution changes the statistic to $d_0 + \delta h$, with

$$\delta \in [\max(-\gamma, -\rho_s),\ \min(\gamma, \rho_t)]. \tag{22}$$

The interval contains zero, giving $R_\gamma \geq d_0$. Maximizing this affine expression at the appropriate endpoint yields

$$R_\gamma(c; \rho) = d_0 + \begin{cases} \min(\gamma, \rho_t)h, & h \geq 0, \\ \min(\gamma, \rho_s)|h|, & h < 0. \end{cases} \tag{23}$$

Both correction terms are bounded above by $\gamma|h|$. This proves $R_\gamma \leq d$, with equality whenever the required extremal redistribution is feasible. In particular, $\rho_s, \rho_t \geq \gamma$ is a sufficient condition. More precisely, equality holds if $h = 0$, or if $h > 0$ and $\rho_t \geq \gamma$, or if $h < 0$ and $\rho_s \geq \gamma$. This completes the proof.

Writing $[x]_+ = \max(x, 0)$ exposes the boundary slack:

$$d - R_\gamma(c; \rho) = [\gamma - \rho_t]_+[h]_+ + [\gamma - \rho_s]_+[-h]_+. \tag{24}$$

Without the simplex constraints, $\max_{|\delta| \leq \gamma}(d_0 + \delta h) = d$ exactly, but some such redistributions may give a negative fusion coefficient. Thus the implemented expression is an upper envelope, not an unconditional equality to a worst-case convex mixture. For normalized non-negative weights, $\frac{1}{2}\|(\rho_s + \delta, \rho_t - \delta) - \rho\|_1 = |\delta|$, establishing the total-variation interpretation of Eq. (6). This local calculation borrows the uncertainty-set viewpoint (**?**); it does not replace the full training objective by a data-distribution minimax problem. The radius is prescribed, not estimated as a confidence bound from samples.

The decomposition

$$d = \frac{c_s + c_t}{2} + \frac{\rho_s - \rho_t}{2}(c_s - c_t) + \gamma|c_s - c_t| \tag{25}$$

separates common loss level, the signed adjustment due to nominal fusion, and the absolute redistribution margin. The final term is a deterministic feature of the same two observations, not an independent third source of evidence. For example, when $c_s = c_t$, the nominal fusion and all feasible reweightings agree and the gap term is zero.

For any fixed prior log-odds b and scale β, let $r_0 = \sigma(b + \beta d_0)$ and $r_\gamma = \sigma(b + \beta[d_0 + \gamma g])$. Then

$$\log \frac{r_\gamma}{1 - r_\gamma} - \log \frac{r_0}{1 - r_0} = \beta\gamma g, \qquad \frac{r_\gamma/(1 - r_\gamma)}{r_0/(1 - r_0)} = e^{\beta\gamma g}. \tag{26}$$

At $b = 0$ and $\beta = 1$ this is Eq. (12). It identifies how the correction changes an operational control distribution; it is not a claim of calibrated odds that the teacher is correct. Since $0 < \sigma'(u) \leq 1/4$,

$$0 \leq r_\gamma - r_0 \leq \frac{\beta\gamma}{4}|c_s - c_t|. \tag{27}$$

At the implemented unit scale, the gap increases w_C and decreases w_A by equal magnitudes bounded by $\Delta_w \gamma |c_s - c_t|/4$. The fixed bridge is unchanged by this correction. In the optional Qwen1.5 $\rightarrow$ GPT2 extension, temperature depends only on c_s and is likewise unaffected by the gap at fixed observations. A small coefficient alone does not bound the effect uniformly over arbitrarily large CE gaps.

For fixed fusion coefficients, perturb the CEs by (δ_s, δ_t). The triangle and reverse-triangle inequalities give

$$\begin{aligned} |\delta d| &\le \rho_s|\delta_s| + \rho_t|\delta_t| + \gamma\big||c_s + \delta_s - c_t - \delta_t| - |c_s - c_t|\big| \\ &\le (\rho_s + \gamma)|\delta_s| + (\rho_t + \gamma)|\delta_t|. \end{aligned} \tag{28}$$

Combining this with the sigmoid derivative yields, for the implemented unit scale,

$$\max\{|\delta w_A|, |\delta w_C|\} \le \frac{\Delta_w}{4}\big[(\rho_s + \gamma)|\delta_s| + (\rho_t + \gamma)|\delta_t|\big]. \tag{29}$$

For the optional Qwen1.5 $\rightarrow$ GPT2 temperature action, the analogous bound is $|\delta\tau| \le (\Delta_\tau/4)|\delta_s|$. This is a pointwise property of the control map conditional on the fusion coefficients, not a stability or convergence theorem for the entire training trajectory.

For $c_s \neq c_t$, the coordinate derivatives of the implemented statistic are

$$\frac{\partial d}{\partial c_s} = \rho_s + \gamma\,\mathrm{sign}(c_s - c_t), \qquad \frac{\partial d}{\partial c_t} = \rho_t - \gamma\,\mathrm{sign}(c_s - c_t). \tag{30}$$

A derivative can be negative when the corresponding fusion coefficient is below γ. The policy is monotone in its fused evidence, not necessarily separately in each CE. At equality the absolute value is nondifferentiable, while the perturbation bounds remain valid. These are derivatives of the mapping, not extra gradients through the detached observations. Equal target CEs do not imply equal non-target distributions, and token-level CE scales vary with segmentation and vocabulary. Consequently, the absolute gap is neither a KL/JS divergence nor a tokenizer-invariant likelihood ratio; convex fusion cannot by itself guarantee scale invariance or teacher-reliability calibration.

A.3 Exact Fusion Gradients and Optimization Scope

For two fusion logits, $\rho_s = \sigma(\ell_s - \ell_t)$ and $\rho_t = 1 - \rho_s$. Holding the detached CE observations fixed gives

$$\frac{\partial d}{\partial \ell_s} = \rho_s\rho_t(c_s - c_t), \qquad \frac{\partial d}{\partial \ell_t} = -\rho_s\rho_t(c_s - c_t). \tag{31}$$

The gap has no direct logit derivative because its two CE arguments and its coefficient are fixed in this partial derivative. In the primary protocol the distillation temperature is fixed, so the three raw channel losses have no additional dependence on the fusion logits during the current forward pass. The optional Qwen1.5 $\rightarrow$ GPT2 temperature action depends on c_s rather than on the fusion logits. With $r = \sigma(d)$,

$$\frac{\partial J_{\mathcal{B}}}{\partial d} = \frac{\lambda_{\mathrm{KD}}\Delta_w}{N_{\mathcal{B}}} r(1 - r)(L_C - L_A). \tag{32}$$

The chain rule proves Eq. (13) exactly for the executed detached-observation objective. The gap changes the update through $r(1 - r)$ even though it has no direct logit derivative. The gradient is zero when $c_s = c_t$ or $L_A = L_C$, and can be small near sigmoid or fusion saturation.

The difference $L_C - L_A$ involves the original valid-token sums; token counts, divergence scales, and channel definitions therefore affect fusion learning. Minimizing this weighted objective can prefer a numerically smaller channel loss and should not be confused with optimizing a held-out objective or identifying which supervision is semantically more reliable. The two trainable logits have only one identifiable difference. They are learned by ordinary gradients, while the binary control distributions are computed analytically conditional on their current values. No posterior distribution over the fusion logits, teacher, or student parameters is inferred.

The guarantees in this appendix are deliberately separated: Proposition 1 bounds local changes of the evidence statistic, Proposition 2 solves a specified control decision, and Eq. (13) describes the gradient used for training. None alone provides an out-of-distribution risk bound, non-convex training convergence guarantee, or guarantee of superiority to a well-tuned static controller.

B Additional Experimental Details And Results

B.1 Evaluation Datasets

Our evaluation follows the instruction-following protocol used by DSKD (Zhang et al., 2024). Databricks-Dolly-15k is used for the distillation data split, while Self-Instruct, Vicuna-Evaluation, Super-NaturalInstructions, and Unnatural Instructions are used as additional held-out evaluation sets. Table 3 summarizes the role and size of each benchmark in our experiments.

Table 3: Evaluation datasets used in our experiments. The test-set sizes and response-length filtering follow the DSKD evaluation protocol (Zhang et al., 2024).

Dataset	Role	Train	Valid	Test	Notes
Dolly	Distillation + evaluation	> 11k	1,000	500	Human-written instruction–response data across multiple task categories (Conover et al., 2023)
SelfInst	Evaluation	–	–	242	Expert-written / self-instruction benchmark for instruction following (Wang et al., 2023)
VicunaEval	Evaluation	–	–	80	Diverse open-ended questions from the Vicuna evaluation suite (Chiang et al., 2023)
S-NI	Evaluation	–	–	1,694	Super-NaturalInstructions subset; responses with length ≥ 11 are retained (Wang et al., 2022)
UnNI	Evaluation	–	–	10,000	Unnatural Instructions subset; responses with length ≥ 11 are retained (Honovich et al., 2023)

Databricks-Dolly-15k. Dolly contains approximately 15k human-generated instruction–response examples and covers capabilities such as brainstorming, classification, closed and open question answering, generation, information extraction, and summarization (Conover et al., 2023). Following the DSKD preprocessing, we use more than 11,000 examples for training, 1,000 for validation, and 500 for testing (Zhang et al., 2024).

Self-Instruct and VicunaEval. Self-Instruct was introduced to study instruction following from model-generated instruction data and includes an expert-written evaluation set for novel tasks (Wang et al., 2023). We use the 242-example SelfInst test set adopted by DSKD. VicunaEval contains 80 diverse open-ended questions spanning multiple categories and is used here as a compact instruction-following evaluation set (Chiang et al., 2023).

S-NI and UnNI. Super-NaturalInstructions contains 1,616 NLP tasks across 76 task types and was designed to test generalization from natural-language task definitions (Wang et al., 2022). We use the 1,694-example S-NI evaluation subset adopted by DSKD after retaining responses with length at least 11. Unnatural Instructions is a large model-generated instruction collection that substantially reduces manual annotation requirements (Honovich et al., 2023); following DSKD, we evaluate on 10,000 examples after the same response-length filtering. These additional sets are used only for evaluation and do not enter the distillation training split.

B.2 Distillation Discrepancy Functions

To examine whether the effectiveness of EAKD depends on a particular distribution-matching objective, we instantiate the dual-space distillation discrepancy with five commonly used KL-family losses: forward Kullback–Leibler divergence (KL), reverse KL (RKL), Jensen–Shannon divergence (JS), skew KL (SKL), and skew reverse KL (SRKL). Similar discrepancy variants have been considered in prior large-language-model distillation studies (Kullback & Leibler, 1951; Lin, 1991; Gu et al., 2024; Ko et al., 2024; Zhang et al., 2024).

Let p denote the teacher distribution and q_θ denote the student distribution over the same aligned output space. In practice, both distributions are obtained from temperature-scaled logits before evaluating the discrepancy. For clarity, we omit the token and conditioning indices below.

Forward KL divergence (KL). The standard forward KL divergence is

$$D_{\mathrm{KL}}(p\|q_\theta) = \sum_i p_i \log \frac{p_i}{q_{\theta,i}}. \tag{33}$$

Forward KL is the conventional distribution-matching objective used in knowledge distillation (Kullback & Leibler, 1951; Hinton et al., 2015). Because the expectation is taken under the teacher distribution, minimizing Eq. (33) encourages the student to allocate probability mass to regions supported by the teacher distribution.

Reverse KL divergence (RKL). Reverse KL exchanges the two arguments:

$$D_{\mathrm{RKL}}(p\|q_\theta) = D_{\mathrm{KL}}(q_\theta\|p) = \sum_i q_{\theta,i} \log \frac{q_{\theta,i}}{p_i}. \tag{34}$$

Compared with forward KL, RKL places the expectation under the student distribution and is commonly associated with a more mode-seeking behavior. MiniLLM adopts reverse KL for generative language-model distillation and argues that it can reduce the tendency of the student to assign excessive probability to low-probability regions of the teacher distribution (Gu et al., 2024).

Jensen–Shannon divergence (JS). JS divergence symmetrizes KL through the midpoint distribution

$$m = \frac{1}{2}(p + q_\theta), \tag{35}$$

and is defined as

$$D_{\mathrm{JS}}(p\|q_\theta) = \frac{1}{2} D_{\mathrm{KL}}(p\|m) + \frac{1}{2} D_{\mathrm{KL}}(q_\theta\|m). \tag{36}$$

Unlike KL and RKL, JS is symmetric with respect to the two distributions and remains finite when their supports differ (Lin, 1991). It therefore provides a balanced alternative to the two directional KL objectives.

Skew KL divergence (SKL). DistiLLM introduces skew KL by replacing the student distribution in the second argument of forward KL with a convex mixture of the teacher and student distributions (Ko et al., 2024). For skew coefficient $\lambda \in [0, 1]$, we define

$$\tilde{q}_\lambda = \lambda p + (1 - \lambda) q_\theta, \tag{37}$$

and

$$D^{(\lambda)}_{\mathrm{SKL}}(p\|q_\theta) = D_{\mathrm{KL}}\left(p\|\lambda p + (1 - \lambda) q_\theta\right). \tag{38}$$

The mixture smooths the comparison distribution by retaining explicit teacher mass, reducing the extremity of the standard forward-KL comparison when the two distributions are poorly matched. Following DistiLLM and DSKD, we use the same skew coefficient as in the corresponding baseline configuration (Ko et al., 2024; Zhang et al., 2024).

Skew reverse KL divergence (SRKL). The reverse counterpart applies the same idea to RKL. We construct

$$\tilde{p}_\lambda = (1 - \lambda) p + \lambda q_\theta, \tag{39}$$

and define

$$D^{(\lambda)}_{\mathrm{SRKL}}(p\|q_\theta) = D_{\mathrm{KL}}\left(q_\theta\|(1 - \lambda) p + \lambda q_\theta\right). \tag{40}$$

SRKL therefore retains the student-weighted expectation of RKL while softening the reference distribution through a teacher–student mixture (Ko et al., 2024). SKL and SRKL were proposed to obtain more stable KL-based objectives for autoregressive language-model distillation, and DSKD adopts both as alternative discrepancy functions in its dual-space evaluation (Ko et al., 2024; Zhang et al., 2024).

These five losses differ only in how the aligned teacher and student distributions are compared. The EAKD controller itself is unchanged: the same mini-batch evidence and adaptive supervision-allocation rule are used for every discrepancy. The corresponding experiments therefore test whether the benefit of adaptive allocation is tied to one particular distribution-matching objective.

B.3 Representation Structure Distance

Directly comparing teacher and student hidden vectors is generally inappropriate when their representation dimensions differ. Following DSKD (Zhang et al., 2024), we instead compare the *relational structure* induced by token-level output hidden states. For a sequence with n aligned token positions and output hidden states $\{h_i\}_{i=1}^n$, we construct a cosine-structure matrix

$$M_{\cos}(i, j) = \frac{h_i^\top h_j}{\|h_i\|_2 \|h_j\|_2} \in \mathbb{R}^{n \times n}, \tag{41}$$

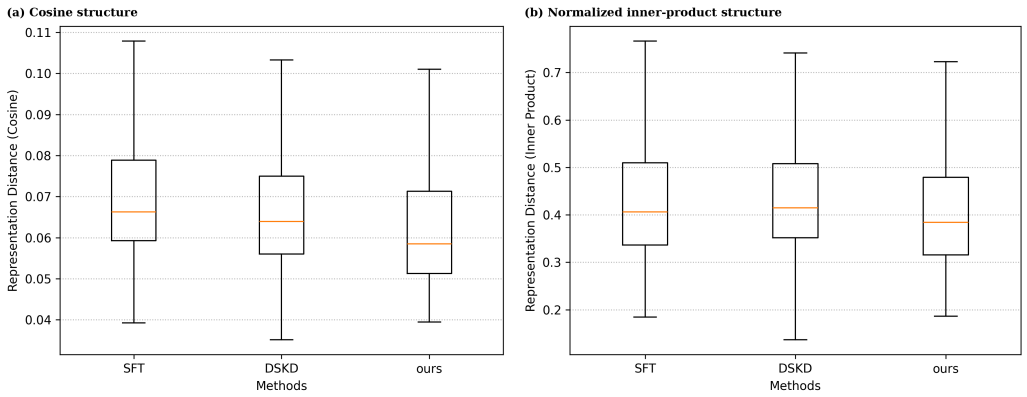


Figure 4: Distribution of teacher–student representation-structure distances for LLaMA2-7B → TinyLLaMA-1.1B on Dolly. We compare the SFT student, DSKD, and EAKD (Ours) under (a) cosine-based relational structure and (b) normalized inner-product structure. Each box summarizes the sample-level structure distances over the evaluation set; lower values indicate that the student's internal relational geometry is closer to that of the teacher. EAKD exhibits a downward-shifted distance distribution under both structure definitions, suggesting improved preservation of teacher-side representation structure.

and a normalized inner-product structure matrix

$$M_{\mathrm{prod}}(i,j) = \frac{h_i^\top h_j}{\sum_{k=1}^{n} h_i^\top h_k} \in \mathbb{R}^{n\times n}. \tag{42}$$

Let $M_\star^t$ and $M_\star^s$ denote the corresponding teacher and student matrices, where $\star \in \{\cos, \mathrm{prod}\}$. We measure their discrepancy with the element-wise L_1 distance

$$D_{\cos} = \sum_{i=1}^{n}\sum_{j=1}^{n} \left| M_{\cos}^t(i,j) - M_{\cos}^s(i,j) \right|, \tag{43}$$

$$D_{\mathrm{prod}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \left| M_{\mathrm{prod}}^t(i,j) - M_{\mathrm{prod}}^s(i,j) \right|. \tag{44}$$

Smaller values indicate that the student's token–token relational geometry is closer to the teacher's. For a set of N examples, we report the distribution of per-example distances and, when needed, their average

$$\overline{D}_\star = \frac{1}{N}\sum_{m=1}^{N} D_\star^{(m)}. \tag{45}$$

For this diagnostic we use the shared-vocabulary LLaMA2-7B → TinyLLaMA-1.1B setting, which provides a direct token-position correspondence and avoids an additional cross-tokenizer alignment step inside the diagnostic itself. Following DSKD, we evaluate 1,000 examples from the Dolly training split. Figure **??** is designed to compare SFT, DSKD, and EAKD using the per-example $D_{\cos}$ and D_{prod} values. Each box summarizes the distribution across examples; lower medians and lower overall distributions indicate closer representation structure to the LLaMA2 teacher.

B.4 GPT-4 Evalution

We complement ROUGE-L with pairwise comparisons on 100 Dolly examples for Mistral-7B → TinyLLaMA-1.1B. The evaluator is recorded as GPT-4-mini in the experimental setup. For each example, EAKD and DSKD responses are randomly assigned to Response A and Response B. The evaluator considers instruction following, helpfulness, relevance, factual correctness, logical coherence, and completeness, and is instructed not to prefer a response merely because it is longer. It returns `A`, `B`, or `Tied`.

Figure 5 reports EAKD-win/tie/DSKD-win percentages of $38/40/22$ for KL, $40/29/31$ for RKL, and $37/28/35$ for JS. EAKD has a higher raw win percentage in all three comparisons, although the JS difference is small. These results are supplementary descriptive evidence from this evaluator and sample, not a claim of statistical significance or comprehensive factual correctness.

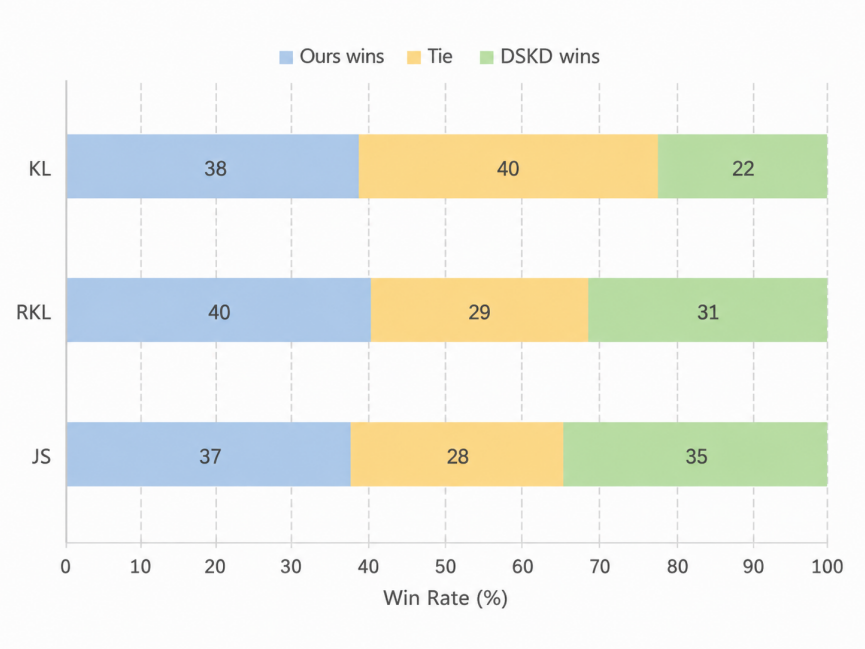


Figure 5: GPT-4-mini-based pairwise evaluation of EAKD and DSKD for Mistral-7B → TinyLLaMA-1.1B on 100 Dolly evaluation examples. KL, reverse KL (RKL), and Jensen–Shannon (JS) objectives are compared. Each bar shows the percentages of EAKD wins, ties, and DSKD wins.

B.5 Hyperparameter Sensitivity

We further examine the sensitivity of EAKD to the two hyperparameters that directly determine the analytical controller: the admissible range of the dynamic loss weights and the coefficient assigned to the student–teacher CE discrepancy. The former controls the range of supervision redistribution, whereas the latter determines how strongly a disagreement between the two CE observations modifies the fused control evidence. Unless otherwise specified, all other training settings follow the primary protocol in Sec. 4.1.

Dynamic Weight-Bound Sensitivity The analytical controller constrains each adaptive coefficient to $[w_{\min}, w_{\max}]$. We fix $w_{\min} = 0.1$ and vary $w_{\max} \in \{5, 10, 15, 20\}$ for both Mistral-7B → TinyLLaMA-1.1B and LLaMA2-7B → TinyLLaMA-1.1B. As summarized in Table 4, $w_{\max} = 10$ gives the highest average ROUGE-L for both teacher–student configurations, reaching 26.90 and 27.13, respectively.

For Mistral-7B → TinyLLaMA-1.1B, reducing the upper bound from the default $w_{\max} = 10$ to 5 lowers the average ROUGE-L from 26.90 to 26.39, while increasing it to 15 and 20 results in averages of 26.54 and 25.47, respectively. The shared-vocabulary LLaMA2-7B → TinyLLaMA-1.1B setting shows a similar overall preference for the intermediate bound: $w_{\max} = 10$ achieves the highest average of 27.13, compared with 26.47, 26.83, and 26.69 for $w_{\max} = 5$, 15, and 20, respectively. The sensitivity is milder in the LLaMA2 setting, where relatively large bounds remain competitive, and individual evaluation sets do not necessarily favor the same value. For example, S-NI and UnNI obtain their highest scores at $w_{\max} = 15$, while VicunaEval peaks at $w_{\max} = 20$. Nevertheless, the aggregate performance of both teacher–student configurations is highest at $w_{\max} = 10$, suggesting that a moderate allocation range provides a better balance than either overly restrictive or excessively large coefficient bounds.

Changing $w_{\max}$ modifies two quantities simultaneously. It enlarges the admissible redistribution between the $T \to S$ CE and $S \to T$ KD channels, but it also changes the absolute coefficient scale because $w_{T2S\text{-}KD} = (w_{\min} + w_{\max})/2$ and $w_{T2S\text{-}CE} + w_{S2T\text{-}KD} = w_{\min} + w_{\max}$. The experiment should therefore be interpreted as sensitivity to the overall controller range rather than as an isolated test of policy flexibility. The degradation at excessively large bounds indicates that simply allowing stronger distillation coefficients is not uniformly beneficial.

Table 4: Sensitivity to the dynamic loss-weight bounds for the two primary TinyLLaMA-1.1B distillation settings. We fix $w_{\min} = 0.1$ and vary $w_{\max}$. Results are ROUGE-L (↑), reported as means with standard deviations in subscript.

Weight Bound	Dolly	SelfInst	VicunaEval	S-NI	UnNI	Avg.
Mistral-7B → TinyLLaMA-1.1B						
$w_{\min} = 0.1,\ w_{\max} = 5$	$27.31_{\pm 0.16}$	$19.59_{\pm 0.68}$	$18.67_{\pm 0.54}$	$33.15_{\pm 0.15}$	$33.24_{\pm 0.08}$	26.39
$w_{\min} = 0.1,\ w_{\max} = 10$	$\mathbf{27.73}_{\pm 0.25}$	$19.61_{\pm 0.51}$	$\mathbf{20.63}_{\pm 0.42}$	$\mathbf{33.45}_{\pm 0.17}$	$33.07_{\pm 0.08}$	**26.90**
$w_{\min} = 0.1,\ w_{\max} = 15$	$27.37_{\pm 0.51}$	$19.61_{\pm 0.72}$	$19.33_{\pm 0.24}$	$32.67_{\pm 0.25}$	$\mathbf{33.71}_{\pm 0.11}$	26.54
$w_{\min} = 0.1,\ w_{\max} = 20$	$23.37_{\pm 0.61}$	$\mathbf{19.85}_{\pm 0.42}$	$19.08_{\pm 0.49}$	$32.76_{\pm 0.28}$	$32.29_{\pm 0.12}$	25.47
LLaMA2-7B → TinyLLaMA-1.1B (Same Vocabulary)						
$w_{\min} = 0.1,\ w_{\max} = 5$	$27.64_{\pm 0.47}$	$19.52_{\pm 0.12}$	$19.35_{\pm 0.52}$	$33.84_{\pm 0.19}$	$32.04_{\pm 0.25}$	26.47
$w_{\min} = 0.1,\ w_{\max} = 10$	$\mathbf{27.93}_{\pm 0.45}$	$\mathbf{20.92}_{\pm 0.32}$	$20.22_{\pm 0.32}$	$33.51_{\pm 0.21}$	$33.07_{\pm 0.11}$	**27.13**
$w_{\min} = 0.1,\ w_{\max} = 15$	$27.47_{\pm 0.39}$	$19.65_{\pm 0.62}$	$19.75_{\pm 0.35}$	$\mathbf{34.01}_{\pm 0.16}$	$\mathbf{33.27}_{\pm 0.21}$	26.83
$w_{\min} = 0.1,\ w_{\max} = 20$	$27.03_{\pm 0.51}$	$20.74_{\pm 0.66}$	$\mathbf{20.34}_{\pm 0.54}$	$32.84_{\pm 0.27}$	$32.51_{\pm 0.07}$	26.69

Gap-Correction Weight Sensitivity The second controller-specific hyperparameter is the gap-correction coefficient γ in the control evidence,

$$d = \rho_s c_s + \rho_t c_t + \gamma |c_s - c_t|, \tag{46}$$

where c_s and c_t denote the student and teacher CEs, respectively. The learned convex fusion captures the overall loss level, while the absolute-gap term provides an additional discrepancy signal when the two observations diverge. We therefore vary $\gamma \in \{0, 0.1, 0.2\}$ while keeping the dynamic loss-weight bounds fixed to the default interval $[0.1, 10]$.

Table 5 shows that a small but nonzero gap correction is beneficial in both primary teacher–student configurations. For Mistral-7B → TinyLLaMA-1.1B, removing the gap term with $\gamma = 0$ yields an average ROUGE-L of 26.15. Introducing the discrepancy correction with $\gamma = 0.1$ increases the average to 26.90, whereas further increasing the coefficient to $\gamma = 0.2$ reduces the average to 26.69. The same pattern is observed for the shared-vocabulary LLaMA2-7B → TinyLLaMA-1.1B setting, where the average improves from 26.47 at $\gamma = 0$ to 27.13 at $\gamma = 0.1$, and then decreases to 26.76 at $\gamma = 0.2$.

Although the best coefficient for an individual evaluation set is not always the same, $\gamma = 0.1$ achieves the highest average performance for both teacher–student pairs. These results indicate that the student–teacher CE gap is useful as a supplementary control signal, but assigning it excessive influence is not necessary. When $\gamma = 0$, the controller relies only on the learned convex fusion of student and teacher CEs; a moderate positive correction improves the aggregate result, while a stronger correction does not provide further benefit. We therefore set $\gamma = 0.1$ in all primary experiments.

B.6 Temperature-Augmented Heterogeneous Stress Test: Qwen1.5-1.8B → GPT2-120M

The primary experiments deliberately use a unified LoRA-based TinyLLaMA protocol and adapt only the loss weights. We additionally evaluate a more heterogeneous setting with a substantially smaller GPT2-120M student and Qwen1.5-1.8B teacher. This experiment differs in two ways: GPT2 and Qwen1.5 use full fine-tuning rather than the LoRA protocol of the main experiments, and the analytical controller additionally adapts the shared distillation temperature. We therefore report it separately rather than pooling it into the primary table.

The GPT2 student is distilled for 20 epochs at learning rate 5×10^{-4} with batch size 8, while the Qwen1.5 teacher is adapted for 10 epochs at 2×10^{-5} with batch size 4. The dynamic directional

Table 5: Sensitivity to the student–teacher CE gap coefficient γ for the two primary TinyLLaMA-1.1B distillation settings. The dynamic loss-weight bounds are fixed to $[0.1, 10]$. Results are ROUGE-L (↑), reported as means with standard deviations in subscript.

γ	Dolly	SelfInst	VicunaEval	S-NI	UnNI	Avg.
Mistral-7B → TinyLLaMA-1.1B						
0	$26.61_{\pm 0.07}$	$\mathbf{19.64}_{\pm 0.33}$	$19.57_{\pm 0.41}$	$32.75_{\pm 0.04}$	$32.19_{\pm 0.11}$	26.15
0.1	$27.73_{\pm 0.25}$	$19.61_{\pm 0.51}$	$\mathbf{20.63}_{\pm 0.42}$	$\mathbf{33.45}_{\pm 0.17}$	$33.07_{\pm 0.08}$	**26.90**
0.2	$27.26_{\pm 0.06}$	$19.07_{\pm 0.94}$	$\mathbf{20.63}_{\pm 0.36}$	$33.37_{\pm 0.04}$	$\mathbf{33.11}_{\pm 0.01}$	26.69
LLaMA2-7B → TinyLLaMA-1.1B (Same Vocabulary)						
0	$27.66_{\pm 0.04}$	$19.09_{\pm 0.14}$	$\mathbf{20.50}_{\pm 0.04}$	$32.91_{\pm 0.02}$	$32.18_{\pm 0.03}$	26.47
0.1	$\mathbf{27.93}_{\pm 0.45}$	$\mathbf{20.92}_{\pm 0.32}$	$20.22_{\pm 0.32}$	$\mathbf{33.51}_{\pm 0.21}$	$33.07_{\pm 0.11}$	**27.13**
0.2	$27.85_{\pm 0.41}$	$19.18_{\pm 0.05}$	$20.23_{\pm 0.14}$	$33.31_{\pm 0.04}$	$\mathbf{33.21}_{\pm 0.07}$	26.76

weights still use Eqs. (5) and (11). In addition, student CE controls a second posterior-mean action,

$$r_\tau = \sigma(c_s), \qquad \tau_{\mathrm{KD}} = \tau_{\mathrm{Tea}} = \tau_{\min} + (\tau_{\max} - \tau_{\min}) r_\tau, \qquad (\tau_{\min}, \tau_{\max}) = (1.0, 4.0). \quad (47)$$

Thus the Qwen1.5 → GPT2 setting uses the same evidence-driven weight redistribution as the primary method plus an additional student-CE-based temperature action.

Table 6: ROUGE-L (↑) for the heterogeneous Qwen1.5-1.8B → GPT2-120M stress test. Unlike the two primary LoRA settings, this configuration uses full fine-tuning and additionally activates the analytical temperature action in Eq. (47). Results are means with standard deviations in subscript.

Methods	Dolly	SelfInst	VicunaEval	S-NI	UnNI	Average
Teacher	$27.19_{\pm 0.23}$	$14.64_{\pm 0.64}$	$16.30_{\pm 0.37}$	$27.55_{\pm 0.30}$	$31.42_{\pm 0.11}$	23.42
SeqKD (Kim & Rush, 2016)	$23.40_{\pm 0.21}$	$9.36_{\pm 0.38}$	$15.37_{\pm 0.35}$	$15.16_{\pm 0.17}$	$17.34_{\pm 0.11}$	16.13
MinED (Wan et al., 2024)	$24.41_{\pm 0.61}$	$10.60_{\pm 0.39}$	$15.86_{\pm 0.42}$	$16.76_{\pm 0.28}$	$19.68_{\pm 0.12}$	17.46
ULD (Boizard et al., 2025)	$23.77_{\pm 0.41}$	$9.67_{\pm 0.50}$	$14.99_{\pm 0.55}$	$17.60_{\pm 0.21}$	$19.49_{\pm 0.12}$	17.11
MultiLevelOT (Cui et al., 2025)	$24.16_{\pm 0.31}$	$10.58_{\pm 0.39}$	$14.90_{\pm 0.38}$	$18.25_{\pm 0.31}$	$22.06_{\pm 0.11}$	17.99
DSKDv2 (Zhang et al., 2025)	$\mathbf{24.90}_{\pm 0.29}$	$11.63_{\pm 0.28}$	$\mathbf{15.84}_{\pm 0.38}$	$\mathbf{21.01}_{\pm 0.34}$	$\mathbf{24.04}_{\pm 0.13}$	**19.48**
DSKD (Zhang et al., 2024)	$24.73_{\pm 0.47}$	$11.15_{\pm 0.34}$	$15.31_{\pm 0.38}$	$17.20_{\pm 0.24}$	$20.57_{\pm 0.08}$	17.79
EAKD + Temp.	$24.84_{\pm 0.29}$	$\mathbf{11.79}_{\pm 0.46}$	$15.41_{\pm 0.29}$	$20.06_{\pm 0.16}$	$22.33_{\pm 0.14}$	18.89 (+1.10↑)

EAKD with the temperature extension improves the DSKD average from 17.79 to 18.89 in this heterogeneous stress test, although DSKDv2 remains stronger in the overall average. We emphasize that temperature adaptation is not part of the primary EAKD protocol. For the two TinyLLaMA-1.1B settings used in the main experiments, namely Mistral-7B → TinyLLaMA-1.1B and LLaMA2-7B → TinyLLaMA-1.1B, we keep the distillation temperature fixed at $\tau = 2.0$. This choice follows the validation-based configuration adopted by DSKD (Zhang et al., 2024), which selects $\tau = 2.0$ for TinyLLaMA-based distillation. More importantly, our preliminary ablations show that introducing batch-wise temperature adaptation on these two model pairs provides no additional gain and can instead slightly degrade performance. This suggests that, once the temperature is already well calibrated, further perturbing the softening scale is unnecessary and may interfere with the effect of adaptive supervision allocation.

We therefore isolate *weight allocation* as the only adaptive component in the primary experiments, so that improvements can be attributed directly to the redistribution of supervision among the T2S-CE, T2S-KD, and S2T-KD channels. The Qwen1.5-1.8B → GPT2-120M setting is qualitatively different: the teacher–student pair is substantially more heterogeneous, and our empirical results indicate that supplementing dynamic weighting with analytical temperature control is beneficial. We consequently retain the temperature extension only for this appendix stress test. Because both the model pair and the controller differ from the primary LoRA protocol, we regard this result as supplementary evidence rather than part of the main comparison.

C Implementation Procedure

Algorithm 1 summarizes the primary mini-batch weight-control procedure. The student, projectors, and two fusion logits are trainable; the teacher remains in evaluation mode with gradients disabled.

The fusion logits belong to the distillation wrapper and are included in its optimizer. CE observations are detached, while the formula-derived weights remain differentiable tensors. The generalized allocation distribution uses a fresh uniform prior and unit evidence scale at each batch; its mean is evaluated directly, without sampling, accumulated posterior odds, or an iterative inner solver. The primary temperature is fixed at $\tau_0 = 2.0$.

Algorithm 1 Mini-batch analytical control for dual-space distillation

Require: Student $\mathcal{S}_\theta$, frozen teacher $\mathcal{T}_\phi$, projectors ξ, trainable fusion logits (ℓ_s, ℓ_t), gap coefficient $\gamma = 0.1$, fixed temperature $\tau_0 = 2.0$, gradient-accumulation length G

for each micro-batch $\mathcal{B}$ **do**
  Run the student; compute logits, hidden states, and L_{CE}
  Run the frozen teacher under no-gradient mode
  Compute per-example native-token CEs and their local batch means
  Detach means; synchronize them across equal-sized data-parallel batches
  $[\rho_s, \rho_t] \leftarrow \mathrm{softmax}([\ell_s, \ell_t])$
  $g \leftarrow |c_s - c_t|$; $d \leftarrow \rho_s c_s + \rho_t c_t + 0.1g$
  $r \leftarrow \sigma(d)$ ▷ Generalized allocation distribution
  Compute (w_A, w_B, w_C) by Eq. (11), keeping their fusion gradient path
  Use the fixed shared temperature τ_0 for both distillation spaces
  Compute cross-model attention and masked channel losses in Eq. (2)
  Form $J_{\mathcal{B}}$ with the driver's shared batch normalizer
  Accumulate gradients using the same G convention for all comparisons
  if the gradient-accumulation boundary is reached **then**
    Update $\theta, \xi, \ell_s, \ell_t$; clear accumulated gradients
  end if
end for

In the primary protocol, the teacher-space forward KL and the student-space discrepancy use the fixed shared temperature $\tau_0 = 2.0$; no additional external τ^2 multiplier is applied in this criterion. The student-space discrepancy otherwise follows the configured divergence implementation. Appendix B.6 adds the temperature action in Eq. (47) before these channel losses are evaluated. Valid-token masks, reliable-token selection, reductions, and the training driver's `batch_denom` are retained. The local mean CE aggregation equals a global sample mean when per-rank batch sizes are equal; unequal batches require sum-and-count aggregation.

The primary analysis trace records per-sample CEs together with the current batch's fusion coefficients, absolute CE gap, and directional weights; the Qwen1.5 $\rightarrow$ GPT2 extension additionally records the dynamic temperature. Repeating the same control values for each sample in a batch is a logging convention rather than reuse of a policy across different batches. The per-sample gather used for tracing is separate from the synchronization of the two controller statistics.