RETACO: RESIDUAL-TARGET CONTROL FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation trains a student with teacher feedback on student-generated prefixes, but communicating or storing full-vocabulary teacher distributions at every generated token is costly. Entropy-aware on-policy distillation (EOPD) supplements reverse Kullback-Leibler (KL) divergence with forward supervision over the teacher's top-k tokens to help recover plausible tokens that the student underestimates. Renormalizing the retained teacher probabilities assigns all target mass to the selected tokens and none to the omitted vocabulary. We prove that the loss continues to push the student's selected mass toward one even after the relative probabilities within the selected set match the teacher's. Consequently, matching the teacher is not a stationary point whenever the omitted vocabulary has positive teacher probability. This finding motivates ReTaCo (Residual-Target Control for On-Policy Distillation), which combines a single-sample estimator whose expectation equals the full-vocabulary reverse KL with a forward target that represents the top-k tokens individually and groups all remaining tokens into one residual symbol. For teacher selected mass m, the residual target is $(1-\beta)(1-m)$, while the relative probabilities of the selected tokens remain unchanged. Thus, $\beta = 0$ preserves teacher mass, and larger values explicitly allocate more target mass to selected tokens. At a fixed prefix, we prove that the population objective has a unique optimum, whose selected mass lies between m and $m + \beta(1-m)$ and increases monotonically with β. At $\beta = 0$, the forward term also retains non-vanishing recovery gradients for underestimated selected tokens. Numerical optimization confirms the predicted optimal mass and conditional distributions. Extensive experiments demonstrate the superiority of ReTaCo, outperforming EOPD on most mathematics and code benchmarks.

1 INTRODUCTION

Transferring the capabilities of large language models to smaller models is an important step toward making these capabilities accessible under limited computational resources. On-policy distillation (OPD) addresses this goal by training a student on its own generated prefixes, with token-level supervision from a stronger teacher (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025). Because the teacher provides feedback on contexts that the student actually encounters, OPD reduces the mismatch between training and inference while building on earlier approaches to knowledge distillation (Hinton et al., 2015; Kim & Rush, 2016). However, transmitting the teacher's full-vocabulary probabilities at every generated token incurs substantial communication costs. Sending only the teacher's top-k tokens and their probabilities reduces this cost, but leaves the training objective to determine how the omitted probability mass should be treated.

Reverse KL can provide a weak signal for recovering teacher-supported tokens that the student rarely samples. Entropy-aware on-policy distillation (EOPD) (Jin et al., 2026) addresses this problem by adding forward KL at positions with high teacher entropy, using renormalized top-k teacher probabilities. This renormalization preserves the relative probabilities of the selected teacher-supported tokens, but it increases their total probability to one. When evaluated against the student's full-vocabulary probabilities, the resulting loss encourages the student to recover the selected teacher-supported tokens while also moving probability away from the remaining vocabulary.

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.

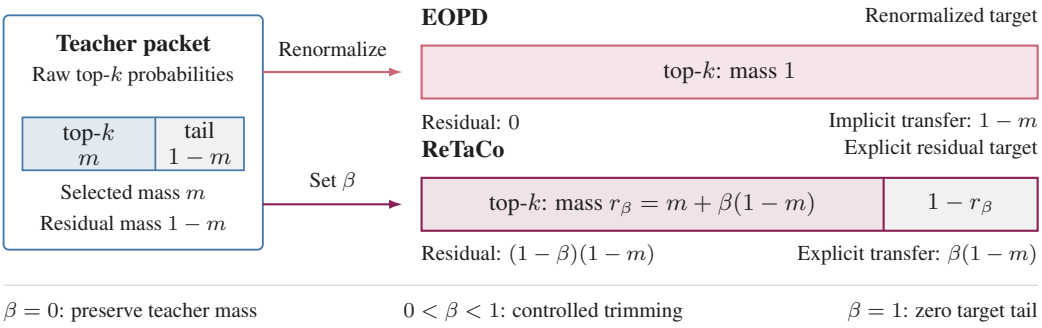


Figure 1: **Residual probability as an explicit forward target.** The teacher packet retains selected mass m and residual mass $1-m$. EOPD renormalizes the selected probabilities to sum to one, assigning zero target mass to the residual. ReTaCo instead sets the residual target to $(1-\beta)(1-m)$ and transfers $\beta(1-m)$ to the selected tokens. Both targets preserve the teacher's conditional proportions within the selected set. Bar widths are schematic.

To make this effect precise, let q and p denote the teacher and student next-token distributions. For the teacher's top-k set $\mathcal{S}$, let $m = q(\mathcal{S})$ and $P = p(\mathcal{S})$ denote the total probabilities that the teacher and student assign to this set. We show that the practical forward term decomposes exactly as

$$\mathcal{L}_{\text{ren}} = D_{\text{KL}}(q^{\mathcal{S}} \| p^{\mathcal{S}}) - \log P, \tag{1}$$

where $q^{\mathcal{S}}$ and $p^{\mathcal{S}}$ are the teacher and student distributions conditioned on the selected set. Even after these conditional distributions match, $-\log P$ continues to push the student's selected mass upward. Consequently, exact teacher matching is not a stationary point of this loss when $m < 1$. Reverse KL moderates this pressure but does not remove it. The entropy gate determines where this loss is applied, while the retained mass determines how much teacher probability the target shifts out of the tail.

This decomposition suggests making residual probability an explicit part of the supervision target. We introduce ReTaCo, short for **Re**sidual-**Ta**rget **Co**ntrol, to specify how much probability the forward target retains outside the selected set. We group the remaining tokens into one residual symbol and assign it probability $(1-\beta)(1-m)$. This gives the selected-mass target

$$r_\beta = m + \beta(1-m), \qquad \beta \in [0,1]. \tag{2}$$

Here, $\beta = 0$ preserves the teacher's mass allocation, while $\beta > 0$ transfers a fraction β of the tail mass to the selected set. The relative probabilities within the selected set remain unchanged. Figure 1 illustrates this allocation.

Our contributions are threefold.

- We show that renormalized top-k forward KL implicitly sets the selected-mass target to one. We establish why exact teacher matching is not stationary and characterize the equilibrium when this loss is combined with reverse KL.
- We introduce ReTaCo with an explicit residual target and $O(k)$ teacher communication. At a fixed prefix, we characterize the unique population optimum, bound its selected mass, and prove that this mass increases monotonically with β. We also establish non-vanishing recovery gradients for underestimated selected teacher-supported tokens at $\beta = 0$.
- We numerically verify the predicted mass and conditional distributions, evaluate downstream performance across three model pairs, and characterize the effects of residual targets and objective components through one-epoch Qwen3 ablations.

2 Preliminaries

2.1 On-Policy Distillation and Selected Support

On-policy token distributions. OPD queries a teacher at prefixes generated by the student (Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025). At such a prefix x, let $q = (q_i)_{i \in \mathcal{V}}$

and $p = (p_i)_{i\in\mathcal{V}}$ be the teacher and student next-token distributions over vocabulary $\mathcal{V}$. We study the ideal per-prefix reverse-KL objective $D_{\mathrm{KL}}(p\|q)$ with x and q fixed. This isolates token-level supervision from rollout sampling, stale policies, and clipping in the practical training loop (Section 5).

Selected mass and conditional shape. Let $\mathcal{S}$ contain the teacher's top-k tokens, and let $\mathcal{T}$ denote its complement in $\mathcal{V}$. The selected teacher and student masses are

$$m = q(\mathcal{S}) = \sum_{i\in\mathcal{S}} q_i, \qquad P = p(\mathcal{S}) = \sum_{i\in\mathcal{S}} p_i. \tag{3}$$

For $i \in \mathcal{S}$ and $j \notin \mathcal{S}$, write

$$q_i^{\mathcal{S}} = \frac{q_i}{m}, \quad p_i^{\mathcal{S}} = \frac{p_i}{P}, \qquad q_j^{\mathcal{T}} = \frac{q_j}{1-m}, \quad p_j^{\mathcal{T}} = \frac{p_j}{1-P}. \tag{4}$$

We consider teacher and student distributions with full support, as produced by finite-logit softmax models, and assume $0 < m, P < 1$ in the derivations. Endpoint identities are interpreted by continuity wherever the corresponding limits are well defined. Let $d_{\mathrm{Ber}}(u\|v) = u\log(u/v) + (1-u)\log[(1-u)/(1-v)]$ denote Bernoulli KL divergence. Conditioning on the selected and residual sets gives

$$D_{\mathrm{KL}}(q\|p) = d_{\mathrm{Ber}}(m\|P) + mD_{\mathrm{KL}}(q^{\mathcal{S}}\|p^{\mathcal{S}}) + (1-m)D_{\mathrm{KL}}(q^{\mathcal{T}}\|p^{\mathcal{T}}), \tag{5}$$

$$D_{\mathrm{KL}}(p\|q) = d_{\mathrm{Ber}}(P\|m) + PD_{\mathrm{KL}}(p^{\mathcal{S}}\|q^{\mathcal{S}}) + (1-P)D_{\mathrm{KL}}(p^{\mathcal{T}}\|q^{\mathcal{T}}). \tag{6}$$

The Bernoulli term measures disagreement in mass allocation, while the remaining terms measure disagreement within each set. This separation lets us distinguish recovering selected teacher-supported tokens from changing their total probability.

2.2 Entropy-Aware On-Policy Distillation

EOPD supplements reverse supervision with forward KL to recover teacher-supported tokens that the student underestimates (Jin et al., 2026). The motivation follows from the ideal reverse-KL logit gradient. For student logits z with $p = \mathrm{softmax}(z)$,

$$\frac{\partial D_{\mathrm{KL}}(p\|q)}{\partial z_j} = p_j\left(\log\frac{p_j}{q_j} - D_{\mathrm{KL}}(p\|q)\right).$$

For a fixed full-support teacher, this gradient vanishes as $p_j \to 0$, whereas the full forward-KL gradient $p_j - q_j$ approaches $-q_j$. Forward supervision can therefore supply a recovery signal even when the student rarely samples a teacher-supported token.

EOPD applies its forward term at high-teacher-entropy positions and uses a top-k teacher packet to limit communication. Let $H(q) = -\sum_{i\in\mathcal{V}} q_i \log q_i$ be teacher entropy and $g_{\mathrm{ent}}(x) = \mathbb{I}[H(q) > \tau]$ the gate at prefix x, where τ is an entropy threshold and $\mathbb{I}$ is the indicator function. With forward weight $\lambda > 0$, the gated token-level contribution is $\lambda g_{\mathrm{ent}}(x)\mathcal{L}_{\mathrm{ren}}$. In the full-vocabulary student-normalization configuration, its forward loss is

$$\mathcal{L}_{\mathrm{ren}}(q, p; \mathcal{S}) = \sum_{i\in\mathcal{S}} q_i^{\mathcal{S}} \log\frac{q_i^{\mathcal{S}}}{p_i}. \tag{7}$$

The teacher probabilities are normalized within $\mathcal{S}$, while the denominator p_i retains full-vocabulary normalization. EOPD combines this gated forward contribution with a proximal policy optimization (PPO)-style reverse update. The gate selects supervised prefixes, while normalization determines the target at each active prefix. If the student is also normalized within $\mathcal{S}$, the loss instead becomes $D_{\mathrm{KL}}(q^{\mathcal{S}}\|p^{\mathcal{S}})$; the mass effect below specifically concerns Equation (7).

3 Method

ReTaCo retains EOPD's recovery signal for selected teacher-supported tokens while making the residual target explicit.

3.1 Analyzing the Renormalization Effect in EOPD

EOPD's forward term supervises selected teacher-supported tokens, but its asymmetric normalization also changes the target mass. Factoring the student probabilities into selected mass and conditional shape exposes this second effect.

Proposition 3.1 (Hidden selected-mass target)**.** *For any teacher-selected set $\mathcal{S}$ with student mass $P > 0$,*

$$\mathcal{L}_{\mathrm{ren}}(q, p; \mathcal{S}) = D_{\mathrm{KL}}(q^{\mathcal{S}} \| p^{\mathcal{S}}) - \log P. \tag{8}$$

Hence the conditional shape is minimized at $p^{\mathcal{S}} = q^{\mathcal{S}}$, while the mass term is minimized at $P = 1$.

Substituting $p_i = P p_i^{\mathcal{S}}$ proves the identity. Equation (7) is therefore forward KL from the full-vocabulary target $(q^{\mathcal{S}}, 0_{\mathcal{T}})$ to the student, where $0_{\mathcal{T}}$ assigns zero probability to every tail token. Even when selected conditional proportions match, $-\log P$ continues to favor larger selected mass. The same effect is visible in the logit gradient.

Corollary 3.2 (Teacher matching is not stationary)**.** *Let $p = \mathrm{softmax}(z)$. The logit gradient of Equation (7) is*

$$\frac{\partial \mathcal{L}_{\mathrm{ren}}}{\partial z_j} = \begin{cases} p_j - q_j^{\mathcal{S}}, & j \in \mathcal{S}, \\ p_j, & j \notin \mathcal{S}. \end{cases} \tag{9}$$

At $p = q$ and $m < 1$, every residual logit has a positive gradient and decreases under direct gradient descent on the logits.

Thus, the entropy gate does not remove the normalization effect at active prefixes. It determines *where* supervision applies, whereas m determines the reassigned mass $1 - m$. In general, entropy does not determine top-k mass, so equal-entropy prefixes need not receive equal mass shifts.

Equilibrium with reverse KL. To isolate whether reverse supervision removes this mass shift, replace the practical reverse update by ideal reverse KL and hold an active prefix fixed. In this per-prefix comparison, $\alpha = \lambda > 0$ because $g_{\mathrm{ent}}(x) = 1$. The idealized joint objective is

$$\mathcal{J}_{\mathrm{ren}}(p) = D_{\mathrm{KL}}(p \| q) + \alpha \mathcal{L}_{\mathrm{ren}}(q, p; \mathcal{S}), \qquad \alpha > 0. \tag{10}$$

For fixed P, all conditional KL terms are minimized at the teacher conditionals. The remaining objective is $d_{\mathrm{Ber}}(P \| m) - \alpha \log P$, so its optimum directly quantifies the competition between mass matching and the unit-mass forward target.

Proposition 3.3 (Mass inflation under the renormalized target)**.** *The selected-mass optimum of Equation (10) is the unique $P^* \in (0, 1)$ satisfying*

$$\log \frac{P^*(1-m)}{m(1-P^*)} = \frac{\alpha}{P^*}. \tag{11}$$

For every $m \in (0, 1)$, $P^* > m$. As $m \to 1$, the residual student mass obeys $1 - P^* = e^{-\alpha}(1-m) + o(1-m)$.

Even ideal reverse KL therefore leaves selected-mass inflation. We next construct a forward target that retains token recovery while explicitly controlling residual mass.

3.2 Residual Representation and Target Construction

Representing residual probability. To specify residual mass without transmitting individual tail probabilities, we group all tokens outside $\mathcal{S}$ into one residual symbol. This symbol belongs only to the supervision representation; the student still predicts over $\mathcal{V}$. Denote the aggregation by $C_{\mathcal{S}}$, which gives

$$C_{\mathcal{S}} p = \big(\{p_i\}_{i \in \mathcal{S}},\, 1 - P\big), \qquad C_{\mathcal{S}} q = \big(\{q_i\}_{i \in \mathcal{S}},\, 1 - m\big). \tag{12}$$

This representation retains selected-token probabilities and aggregate tail mass. By the reverse-KL chain rule,

$$D_{\mathrm{KL}}(p \| q) = D_{\mathrm{KL}}(C_{\mathcal{S}} p \| C_{\mathcal{S}} q) + (1 - P) D_{\mathrm{KL}}(p^{\mathcal{T}} \| q^{\mathcal{T}}), \tag{13}$$

$$D_{\mathrm{KL}}(C_{\mathcal{S}} p \| C_{\mathcal{S}} q) = d_{\mathrm{Ber}}(P \| m) + P D_{\mathrm{KL}}(p^{\mathcal{S}} \| q^{\mathcal{S}}). \tag{14}$$

Aggregation preserves the mass of the tail but omits its conditional divergence. We therefore use the coarse representation for forward supervision and retain sensitivity to individual tail tokens through sampled reverse KL (Section 3.3).

Controlling the residual target. Choose $\beta \in [0,1]$ and retain the fraction $1-\beta$ of teacher residual mass, so that $1 - r_\beta = (1-\beta)(1-m)$. Normalization fixes selected mass at r_β; preserving the teacher conditional proportions then gives

$$r_\beta = m + \beta(1-m), \qquad t_\beta = \left(\left\{ r_\beta \frac{q_i}{m} \right\}_{i \in \mathcal{S}}, 1 - r_\beta \right), \quad \beta \in [0,1]. \tag{15}$$

This target preserves the selected conditional proportions and assigns an explicit probability to the residual. Its forward KL decomposes as

$$D_{\mathrm{KL}}(t_\beta \| C_{\mathcal{S}} p) = d_{\mathrm{Ber}}(r_\beta \| P) + r_\beta D_{\mathrm{KL}}(q^{\mathcal{S}} \| p^{\mathcal{S}}). \tag{16}$$

The Bernoulli term supervises selected mass, and the conditional term supervises the relative probabilities of selected teacher-supported tokens. At $\beta = 0$, $t_\beta = C_{\mathcal{S}} q$ preserves teacher mass; positive β transfers $\beta(1-m)$ to the selected tokens, with zero residual target at $\beta = 1$. Changing β also changes the conditional coefficient r_β, coupling mass supervision with conditional weighting.

3.3 Reverse-KL Estimation and the Combined Objective

The residual target supervises aggregate tail mass but leaves individual tail proportions unspecified. We complement it with a single-sample estimator whose expectation equals the full-vocabulary reverse KL. The estimator uses the teacher probability of a student-sampled token that may lie outside $\mathcal{S}$.

Sampled reverse KL. Let $y \sim p$ be the rollout token and $a_y = \log p_y - \log q_y$. With $\mathrm{sg}(\cdot)$ denoting stop-gradient, we use the following straight-through estimator with the standard low-variance forward value.

$$\widehat{D}^{\mathrm{k3+}}_{\mathrm{RKL}}(p\|q; y) = \mathrm{sg}\big(e^{-a_y} + a_y - 1\big) + \frac{a_y^2}{2} - \mathrm{sg}\left(\frac{a_y^2}{2}\right). \tag{17}$$

The estimator evaluates to $e^{-a_y} + a_y - 1$ in the forward pass and uses the derivative of $a_y^2/2$ in the backward pass. Let θ denote the student parameters. At a fixed prefix with a fixed teacher, sampling from the current student without numerical clipping gives

$$\mathbb{E}_{y\sim p}\left[\widehat{D}^{\mathrm{k3+}}_{\mathrm{RKL}}(p\|q; y)\right] = D_{\mathrm{KL}}(p\|q), \qquad \mathbb{E}_{y\sim p}\left[\nabla_\theta \widehat{D}^{\mathrm{k3+}}_{\mathrm{RKL}}(p\|q; y)\right] = \nabla_\theta D_{\mathrm{KL}}(p\|q). \tag{18}$$

The sampled token is detached during differentiation, and the estimator requires only its teacher log-probability. Neither expectation differentiates through prefix sampling.

Objective. At each on-policy token, ReTaCo uses

$$\mathcal{L}_{\text{sampled-ReTaCo}}(y) = \widehat{D}^{\mathrm{k3+}}_{\mathrm{RKL}}(p\|q; y) + \alpha D_{\mathrm{KL}}(t_\beta \| C_{\mathcal{S}} p) \tag{19}$$

with forward weight $\alpha \geq 0$. The default training loss averages this objective over all unmasked response tokens without an entropy gate. Taking the expectation over a token sampled from the current student gives the ideal unclipped per-prefix population objective

$$\mathcal{J}_{\mathrm{ReTaCo}}(p) = D_{\mathrm{KL}}(p\|q) + \alpha D_{\mathrm{KL}}(t_\beta \| C_{\mathcal{S}} p). \tag{20}$$

At $\beta = 0$, the forward target preserves teacher mass and selected conditional shape, and the reverse estimator recovers full-vocabulary KL in expectation. Positive β specifies deliberate tail trimming. At $\beta = 1$, the forward term equals Equation (7); this endpoint recovers EOPD's forward target at active prefixes, not its entropy gate or PPO updates. When $k = |\mathcal{V}|$, the residual vanishes and Equation (20) reduces to full-vocabulary reverse KL plus α times forward KL.

3.4 Population Optimum and Token Recovery

We now determine how the forward target changes the optimum of the combined objective. For fixed selected mass P, the shape-dependent terms in Equation (20) are

$$P D_{\mathrm{KL}}(p^{\mathcal{S}} \| q^{\mathcal{S}}) + (1-P) D_{\mathrm{KL}}(p^{\mathcal{T}} \| q^{\mathcal{T}}) + \alpha r_\beta D_{\mathrm{KL}}(q^{\mathcal{S}} \| p^{\mathcal{S}}), \tag{21}$$

All three terms vanish simultaneously only at $p^{\mathcal{S}} = q^{\mathcal{S}}$ and $p^{\mathcal{T}} = q^{\mathcal{T}}$. The remaining scalar objective is $g(P) = d_{\mathrm{Ber}}(P\|m) + \alpha d_{\mathrm{Ber}}(r_\beta \| P)$. Thus, the target mass r_β and the optimum mass P^* need not coincide.

Theorem 3.4 (Unique and controllable selected-mass optimum). *Let $m \in (0,1)$, $\alpha > 0$, and $\beta \in [0,1]$. The population objective in Equation (20) has a unique full-support minimizer. Its conditional shapes are $p^{\mathcal{S}} = q^{\mathcal{S}}$ and $p^{\mathcal{T}} = q^{\mathcal{T}}$, and its selected mass P^* is the unique root of*

$$F(P) = \log \frac{P(1-m)}{m(1-P)} + \alpha \frac{P - r_\beta}{P(1-P)} = 0. \tag{22}$$

Moreover,

$$m \le P^* \le r_\beta, \qquad \frac{\mathrm{d}P^*}{\mathrm{d}\beta} > 0 \quad \text{for } \beta \in (0,1), \tag{23}$$

with positive one-sided derivatives at the endpoints. For $\beta = 0$, $P^* = m$ exactly, whereas for $\beta > 0$, both inequalities are strict.

Reverse KL favors teacher mass m, while forward KL favors r_β; their balance places the optimum between these two values. The stationarity function is strictly increasing because

$$F'(P) = \frac{1}{P(1-P)} + \alpha \frac{(P - r_\beta)^2 + r_\beta(1 - r_\beta)}{P^2(1-P)^2} > 0. \tag{24}$$

Together with the endpoint signs of F, this establishes a unique root. Complete proofs appear in Appendix B.

Token recovery at the mass-preserving endpoint. Having characterized the optimum, we now examine the recovery signal for underestimated selected tokens at the mass-preserving endpoint ($\beta = 0$).

Proposition 3.5 (Forward logit gradient). *For student logits z with $p = \mathrm{softmax}(z)$,*

$$\frac{\partial D_{\mathrm{KL}}(t_\beta \| C_{\mathcal{S}} p)}{\partial z_j} = \begin{cases} p_j - r_\beta q_j^{\mathcal{S}}, & j \in \mathcal{S}, \\ p_j \dfrac{r_\beta - P}{1 - P}, & j \notin \mathcal{S}. \end{cases} \tag{25}$$

At $\beta = 0$ and $p_j \to 0$ for a selected token, the gradient approaches $-q_j$, while the target selected mass remains m.

Hence, recovering selected teacher-supported tokens does not require a unit selected-mass target. For $\beta > 0$, the deliberate departure from the aggregated teacher is quantified by total variation distance, denoted by TV,

$$\mathrm{TV}(t_\beta, C_{\mathcal{S}} q) = \beta(1 - m). \tag{26}$$

The transferred mass $\delta_\beta = \beta(1-m)$ measures target displacement. A fixed β has a small absolute effect when teacher top-k mass is near one and a larger effect when less mass is retained.

3.5 Implementation

At each unmasked response position, the student rollout supplies y, and the teacher transmits $\log q_y$ with the top-k token IDs and their *raw* log-probabilities. Full-vocabulary normalization makes their probabilities sum to m, rather than one. We compute $\log m$ by log-sum-exp and gather the corresponding student log-probabilities to obtain P. These quantities define t_β and $C_{\mathcal{S}} p$ for the forward loss; $\log p_y$ and $\log q_y$ give the sampled reverse term in Equation (17). The sampled token, selected set, and all teacher quantities are detached.

The teacher packet contains $O(k)$ values per token. Given full-vocabulary student log-probabilities, the additional loss computation is also $O(k)$. This reduces teacher–student communication; the teacher still computes full-vocabulary softmax normalization and top-k selection. Numerical details appear in Appendix C.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. MATH500, the English text-only mathematics subset of OlympiadBench, AMC (Li et al., 2024), AIME24, and AIME25 assess mathematical reasoning (Hendrycks

Table 1: **Per-benchmark performance (%). Mathematics reports avg@8, code reports Pass@1, and OOD reports accuracy.** OOD scores use the mathematics-distilled checkpoints. Bold marks the best score within each pair.

Method	Mathematics					Code		OOD		
	MATH 500	Olympiad Bench	AMC	AIME 24	AIME 25	Human Eval+	MBPP+	ARC -C	MMLU -Pro	GPQA Diamond
Student: Qwen3-1.7B Teacher: Qwen3-30B-A3B-Instruct-2507										
Sampled OPD	82.40	48.67	52.11	25.42	**22.08**	64.02	52.38	**82.51**	45.79	**18.94**
EOPD	79.80	45.54	48.04	21.67	17.92	66.46	51.59	79.84	42.91	12.44
ReTaCo	**83.78**	**49.65**	**52.56**	**28.75**	21.67	**68.29**	**53.17**	82.37	**46.67**	17.49
Student: Qwen3.5-2B Teacher: Qwen3.5-27B										
Sampled OPD	71.73	40.07	45.63	21.25	17.50	48.17	47.88	88.84	52.61	26.89
EOPD	70.93	41.52	46.08	20.00	17.50	50.61	44.18	89.00	53.73	29.17
ReTaCo	**79.85**	**46.93**	**52.71**	**22.92**	**21.25**	**54.88**	**48.68**	**89.40**	**54.20**	**30.43**
Student: Gemma4-E2B Teacher: Gemma4-26B-A4B										
Sampled OPD	48.18	26.83	31.93	10.00	14.17	75.61	66.38	62.18	28.99	13.26
EOPD	62.50	33.96	34.19	14.17	16.67	74.39	67.20	**86.89**	**43.45**	13.64
ReTaCo	**79.47**	**47.93**	**56.17**	**22.50**	**19.58**	**76.83**	**67.79**	84.84	41.60	**31.00**

et al., 2021; Lightman et al., 2023; He et al., 2024). HumanEval+ and MBPP+ assess code generation through the extended EvalPlus tests (Liu et al., 2023; EvalPlus Team, 2023). ARC-C, MMLU-Pro, and GPQA-Diamond assess out-of-distribution (OOD) performance after mathematics distillation (Clark et al., 2018; Wang et al., 2024; Rein et al., 2023). Here, OOD is relative to mathematics distillation training, not model pretraining. We report avg@8 for mathematics, Pass@1 for code, and accuracy on the OOD benchmarks using the protocols in Appendix F.1.

Models and comparisons. Table 1 lists three teacher–student pairs. Sampled OPD and ReTaCo use the same direct $k3+$ reverse estimator. EOPD uses entropy-gated renormalized forward supervision and PPO-style reverse updates, so comparison with EOPD measures the complete training procedures rather than the residual target alone.

Training and evaluation. Mathematics and code distillation use DAPO-Math-17k and the code subset of Eurus-2-RL-Data, respectively. Default ReTaCo uses $k = 16$, $\alpha = 1$, and $\beta = 0$, with ungated response-token averaging. Mathematics and OOD evaluations use the same epoch-1 checkpoints from mathematics distillation; code evaluations use separately code-distilled checkpoints. Mathematical avg@8 averages correctness over eight completions per problem. Optimization and decoding settings appear in Appendix F.

Downstream performance. ReTaCo achieves higher point estimates than both baselines on all five mathematics tasks for Qwen3.5-2B and Gemma4-E2B, and on four for Qwen3-1.7B (Table 1). For example, Qwen3.5-2B improves from 71.73% to 79.85% on MATH500 relative to Sampled OPD. Sampled OPD remains slightly higher on Qwen3-1.7B AIME25. Across all three pairs, ReTaCo also leads both baselines on HumanEval+ and MBPP+.

Out-of-distribution performance. OOD results are less uniform. For Qwen3.5-2B, ReTaCo leads on all three tasks. Qwen3-1.7B improves on MMLU-Pro but remains below Sampled OPD on ARC-C and GPQA-Diamond. For Gemma4-E2B, ReTaCo leads on GPQA-Diamond, whereas EOPD leads on ARC-C and MMLU-Pro.

4.2 Residual Targets and Component Contributions

Effect of the residual target. Table 2 compares residual targets on Qwen3-1.7B under a common one-epoch training schedule. Adding the default forward term improves four of the five mathematical

Table 2: **Qwen3-1.7B target and component comparisons (avg@8, %).** Let R denote sampled reverse KL, $B_\beta = d_{\mathrm{Ber}}(r_\beta \| P)$, and $C = D_{\mathrm{KL}}(q^{\mathcal{S}} \| p^{\mathcal{S}})$. The default is $R + B_0 + mC$. All variants train for one epoch from the same initial student.

Setting	Objective	MATH 500	Olympiad Bench	AMC	AIME24	AIME25
Forward supervision and residual-target choice						
Full ReTaCo	$R + B_0 + mC$	83.78	49.65	52.56	28.75	21.67
No forward term	R	82.40	48.67	52.11	25.42	22.08
Partial trimming ($\beta = 0.5$)	$R + B_{0.5} + r_{0.5}C$	78.88	43.61	47.74	24.17	21.67
Complete trimming ($\beta = 1$)	$R + B_1 + C$	78.60	43.78	48.95	23.75	21.67
Component variants under one-epoch training						
Mass-only forward term	$R + B_0$	81.65	47.15	50.75	21.67	22.08
Conditional-only forward term	$R + mC$	83.55	49.13	54.37	27.92	21.25
No reverse term	$B_0 + mC$	82.10	47.61	51.05	25.42	21.67

scores over reverse-only supervision, including MATH500 from 82.40% to 83.78%, while AIME25 decreases from 22.08% to 21.67%. Preserving teacher mass ($\beta = 0$) also outperforms partial or complete trimming on MATH500, OlympiadBench, and AMC. On AIME24, trimming reduces accuracy by 4.58 or 5.00 percentage points, while AIME25 is unchanged (Figure 3b). The benefit of residual trimming is therefore task-dependent. Because β changes both the mass target and conditional weight r_β, this comparison evaluates the target as a whole. The $\beta = 1$ row shares EOPD's forward target, while retaining the ungated direct reverse objective used by the other ablations.

Mass and conditional supervision. Among the three component variants in Table 2, conditional-only forward supervision achieves the highest scores on MATH500, OlympiadBench, AMC, and AIME24, while mass-only supervision leads on AIME25. Relative to the full objective, removing the mass term lowers MATH500 from 83.78% to 83.55% and AIME24 from 28.75% to 27.92%, but raises AMC from 52.56% to 54.37%. Thus, the mass constraint does not uniformly improve downstream accuracy, and conditional recovery of selected teacher-supported tokens explains most accuracy gains on these tasks.

Selected and residual probability mass. Figure 2 compares EOPD and ReTaCo on Qwen3-1.7B with the Qwen3-30B-A3B-Instruct-2507 teacher and $k = 16$, over their common interval of steps 1–202. Over the last 20 steps, mean student selected mass is $P = 0.991252$ for ReTaCo and $P = 0.989900$ for EOPD, while teacher mass is $m \approx 0.9977$. Thus, ReTaCo brings the student's selected mass closer to the teacher. The tail view shows the same improvement: ReTaCo assigns less excess probability outside the teacher's top-k set than EOPD. Appendix F.2 complements this comparison with policy-entropy trajectories for Sampled OPD and ReTaCo.

Numerical validation of the population objective. To test the fixed-prefix predictions directly, we next examine the population objective. The equilibria in Figure 3(a) lie between m and r_β, with $P^* = m$ at $\beta = 0$ (Theorem 3.4). For a uniform 80-token teacher with $k = 16$ and $\alpha = 1$, the renormalized target instead gives $P^* = 0.5821$ despite $m = 0.2$. Separately, joint optimization of a categorical student recovers the predicted mass and teacher conditionals, with maximum selected-conditional ℓ_1 error 4.13×10^{-8} (Appendix D).

5 Discussion and Limitations

At $\beta = 0$, the forward objective retains aggregate teacher mass and supervises individual selected teacher-supported tokens, but it cannot resolve the relative probabilities of tail tokens. Increasing k refines this supervision at higher communication cost. Sampled reverse KL retains tail sensitivity in expectation. The equilibrium theorem holds at a fixed prefix under the population objective; it does not guarantee convergence with evolving prefixes, stale policies, or clipped updates.

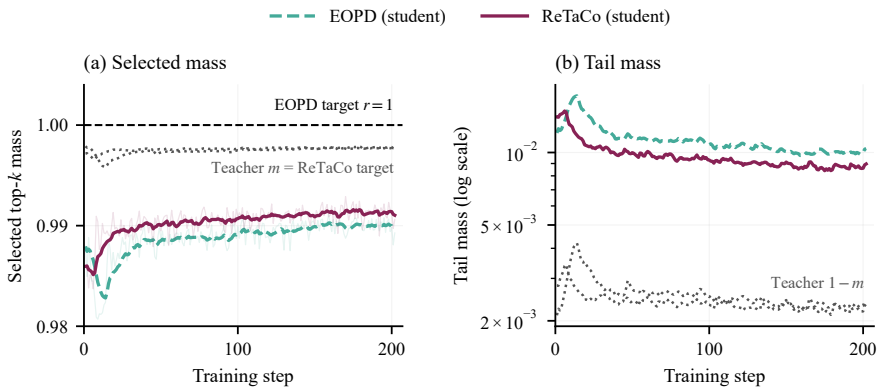


Figure 2: **Selected and residual mass during training.** Qwen3-1.7B student, Qwen3-30B-A3B-Instruct-2507 teacher, and $k = 16$; EOPD and ReTaCo share the displayed interval of steps 1–202. (a) Student mass P on the teacher's top-k set for EOPD (teal, dashed) and ReTaCo (wine, solid). Gray dotted curves show teacher mass m on each method's own rollouts; m is also the ReTaCo forward target at $\beta = 0$. The black dashed line marks EOPD's forward target $r = 1$. (b) The same data expressed as tail mass $1 - P$ and $1 - m$, with a logarithmic vertical axis. Curves use exponential smoothing with span 11; faint traces in (a) show raw values.

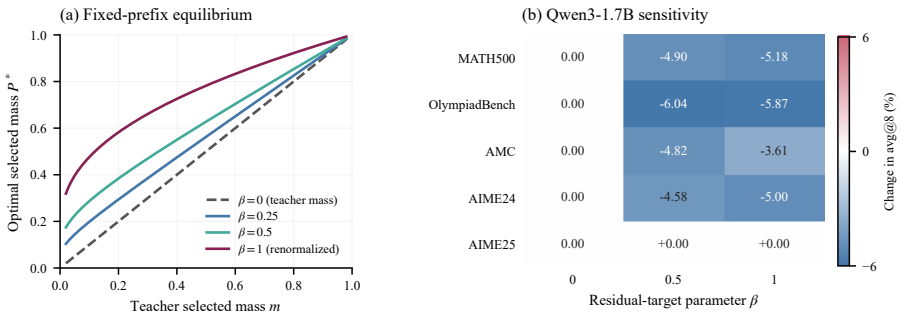


Figure 3: **Residual-target control in analysis and experiments.** (a) Selected-mass equilibria from Equation (22) at $\alpha = 1$; the dashed line denotes teacher mass. (b) Changes in Qwen3-1.7B avg@8 relative to $\beta = 0$, computed from the residual-target rows in Table 2. Positive values indicate higher scores with trimming. All variants train for one epoch.

The mathematics ablations show task-dependent effects of residual trimming and stronger forward supervision. OOD evaluations use mathematics-distilled models and are defined relative to mathematics distillation training, with no claim of exclusion from the models' pretraining corpora. The response-length experiment in Appendix E uses an entropy-gated forward term and a PPO-style reverse objective, so it characterizes that variant's generation behavior.

6 Conclusion

ReTaCo treats residual probability as an explicit supervision target. Its top-k-plus-residual forward target preserves selected conditional proportions, while β controls residual retention and the single-sample reverse estimator recovers full-vocabulary KL in expectation. The fixed-prefix population objective has a unique optimum whose selected mass increases monotonically with β; the mass-preserving endpoint retains recovery gradients for selected teacher-supported tokens. Numerical optimization matches these predictions. Experiments across three teacher–student pairs show higher point estimates on most mathematics and code benchmarks, with improvements across all three OOD benchmarks for Qwen3.5-2B. One-epoch component ablations indicate that conditional recovery explains most accuracy gains, while mass control has task-dependent effects.

AI USE STATEMENT

Generative AI tools assisted with language polishing and editing of the manuscript and with figure preparation. The authors are responsible for the technical content, experimental results, references, and the final manuscript.

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 *International Conference on Learning Representations*, 2024.

Ibtihel Amara, Nazanin Sepahvand, Brett H. Meyer, Warren J. Gross, and James J. Clark. BD-KD: Balancing the divergences for online knowledge distillation. *arXiv preprint arXiv:2212.12965*, 2022.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Sayantan Dasgupta, Trevor Cohn, and Timothy Baldwin. Don't ignore the tail: Decoupling top-k probabilities for efficient language model distillation. *arXiv preprint arXiv:2602.20816*, 2026.

EvalPlus Team. EvalPlus v0.2.0. Official software release, 2023. URL https://github.com/evalplus/evalplus/releases/tag/v0.2.0. Release introducing MBPP+.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

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.

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

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. ToDi: Token-wise distillation via fine-grained divergence control. In *Conference on Empirical Methods in Natural Language Processing*, 2025.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath. GitHub repository, 2024. URL https://github.com/project-numina/aimo-progress-prize.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. arXiv preprint arXiv:2305.01210, 2023. URL https://arxiv.org/abs/2305.01210.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. URL https://thinkingmachines.ai/blog/on-policy-distillation/.

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. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. arXiv preprint arXiv:2311.12022, 2023. URL https://arxiv.org/abs/2311.12022.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024.

KaShun Shum, Minrui Xu, Jianshu Zhang, Zixin Chen, Shizhe Diao, Hanze Dong, Jipeng Zhang, and Muhammad Omer Raza. FIRST: Teach a reliable large language model through efficient trustworthy distillation. *arXiv preprint arXiv:2408.12168*, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574, 2024. URL https://arxiv.org/abs/2406.01574.

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 *International Conference on Computational Linguistics*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

A Complete Derivations

A.1 KL chain rules

Under the assumptions in Section 2, substituting $q_i = mq_i^{\mathcal{S}}$ and $p_i = Pp_i^{\mathcal{S}}$ on the selected set yields

$$\sum_{i\in\mathcal{S}} q_i \log \frac{q_i}{p_i} = \sum_{i\in\mathcal{S}} mq_i^{\mathcal{S}} \left(\log \frac{m}{P} + \log \frac{q_i^{\mathcal{S}}}{p_i^{\mathcal{S}}}\right)$$
$$= m \log \frac{m}{P} + mD_{\mathrm{KL}}(q^{\mathcal{S}} \| p^{\mathcal{S}}). \tag{27}$$

Applying the same expansion to the residual set gives

$$(1-m) \log \frac{1-m}{1-P} + (1-m) D_{\mathrm{KL}}(q^{\mathcal{T}} \| p^{\mathcal{T}}). \tag{28}$$

Adding both contributions establishes Equation (5), and exchanging p and q gives Equation (6). Residual aggregation removes the conditional tail term but retains its mass, as expressed in Equations (13) and (14).

A.2 The sampled reverse-KL estimator

Hold the prefix and teacher distribution fixed, and sample the token from the current student. For $y \sim p$ and $a_y = \log p_y - \log q_y$, the forward value of Equation (17) satisfies

$$\mathbb{E}_{y\sim p}\left[e^{-a_y} + a_y - 1\right] = \sum_{i\in\mathcal{V}} p_i \left(\frac{q_i}{p_i} + \log \frac{p_i}{q_i} - 1\right) \tag{29}$$
$$= D_{\mathrm{KL}}(p\|q). \tag{30}$$

With the sampled token detached, the expected straight-through gradient is

$$\mathbb{E}_{y\sim p}\left[\nabla_\theta \frac{a_y^2}{2}\right] = \sum_{i\in\mathcal{V}} p_i \log \frac{p_i}{q_i} \nabla_\theta \log p_i \tag{31}$$
$$= \nabla_\theta D_{\mathrm{KL}}(p\|q), \tag{32}$$

where the last equality uses $\sum_i p_i \nabla_\theta \log p_i = \nabla_\theta \sum_i p_i = 0$. This proves Equation (18) for the ideal unclipped per-prefix estimator, without differentiation through prefix sampling.

A.3 The implicit target of renormalized forward KL

Proof of Proposition 3.1. Within $\mathcal{S}$, the factorization $p_i = Pp_i^{\mathcal{S}}$ gives

$$\mathcal{L}_{\mathrm{ren}} = \sum_{i\in\mathcal{S}} q_i^{\mathcal{S}} \log \frac{q_i^{\mathcal{S}}}{Pp_i^{\mathcal{S}}}$$
$$= \sum_{i\in\mathcal{S}} q_i^{\mathcal{S}} \log \frac{q_i^{\mathcal{S}}}{p_i^{\mathcal{S}}} - \log P = D_{\mathrm{KL}}(q^{\mathcal{S}} \| p^{\mathcal{S}}) - \log P. \tag{33}$$

The conditional KL vanishes at teacher matching, whereas $-\log P$ decreases strictly on $(0, 1]$ and is minimized at unit selected mass. □

Proof of Corollary 3.2. Ignoring teacher-only constants,

$$\mathcal{L}_{\mathrm{ren}} = -\sum_{i\in\mathcal{S}} q_i^{\mathcal{S}} \log p_i + \mathrm{const}. \tag{34}$$

For $p = \mathrm{softmax}(z)$, $\partial \log p_i / \partial z_j = \mathbb{I}[i = j] - p_j$, where $\mathbb{I}$ is the indicator function. Hence

$$\frac{\partial \mathcal{L}_{\mathrm{ren}}}{\partial z_j} = -\sum_{i\in\mathcal{S}} q_i^{\mathcal{S}} \left(\mathbb{I}[i=j] - p_j\right)$$
$$= \begin{cases} p_j - q_j^{\mathcal{S}}, & j \in \mathcal{S}, \\ p_j, & j \notin \mathcal{S}. \end{cases} \tag{35}$$

At $p = q$, every $j \notin \mathcal{S}$ has logit gradient $q_j > 0$ under the full-support assumption. Teacher matching therefore fails to be stationary whenever residual mass is positive. □

A.4 Equilibrium under the implicit unit-mass target

Proof of Proposition 3.3. By Equations (6) and (8), the conditional terms in Equation (10) are

$$PD_{\mathrm{KL}}(p^{\mathcal{S}}\|q^{\mathcal{S}}) + \alpha D_{\mathrm{KL}}(q^{\mathcal{S}}\|p^{\mathcal{S}}) + (1-P)D_{\mathrm{KL}}(p^{\mathcal{T}}\|q^{\mathcal{T}}). \tag{36}$$

For any fixed $P \in (0,1)$, they have the unique minimizer $p^{\mathcal{S}} = q^{\mathcal{S}}$ and $p^{\mathcal{T}} = q^{\mathcal{T}}$. The remaining scalar objective is

$$g_{\mathrm{ren}}(P) = d_{\mathrm{Ber}}(P\|m) - \alpha \log P. \tag{37}$$

Its first two derivatives are

$$g'_{\mathrm{ren}}(P) = \log \frac{P(1-m)}{m(1-P)} - \frac{\alpha}{P}, \tag{38}$$

$$g''_{\mathrm{ren}}(P) = \frac{1}{P(1-P)} + \frac{\alpha}{P^2} > 0. \tag{39}$$

Because the derivative is strictly increasing and tends to $-\infty$ as $P \downarrow 0$ and to $+\infty$ as $P \uparrow 1$, it has exactly one root. Since $g'_{\mathrm{ren}}(m) = -\alpha/m < 0$, this root satisfies $P^* > m$.

For the asymptotic statement, let $\delta = 1 - m$ and $\epsilon = 1 - P^*$. Equation (11) becomes

$$\log \frac{(1-\epsilon)\delta}{(1-\delta)\epsilon} = \frac{\alpha}{1-\epsilon}. \tag{40}$$

Since the student residual is smaller than the teacher residual, both δ and ϵ approach zero as $m \to 1$. Expansion gives $\log(\delta/\epsilon) = \alpha + o(1)$, and therefore $\epsilon = e^{-\alpha}\delta + o(\delta)$. □

B Proofs for ReTaCo

B.1 Unique population optimum and monotonic control

Proof of Theorem 3.4. Expanding the ideal population objective in Equation (20) gives

$$\begin{aligned}\mathcal{J}_{\mathrm{ReTaCo}} = {} & d_{\mathrm{Ber}}(P\|m) + \alpha d_{\mathrm{Ber}}(r_\beta\|P) \\ & + PD_{\mathrm{KL}}(p^{\mathcal{S}}\|q^{\mathcal{S}}) + (1-P)D_{\mathrm{KL}}(p^{\mathcal{T}}\|q^{\mathcal{T}}) + \alpha r_\beta D_{\mathrm{KL}}(q^{\mathcal{S}}\|p^{\mathcal{S}}).\end{aligned} \tag{41}$$

For fixed interior mass P, the conditional KL terms are non-negative and vanish simultaneously only at $p^{\mathcal{S}} = q^{\mathcal{S}}$ and $p^{\mathcal{T}} = q^{\mathcal{T}}$. Optimizing the conditional shapes therefore reduces the population problem to

$$g(P) = d_{\mathrm{Ber}}(P\|m) + \alpha d_{\mathrm{Ber}}(r_\beta\|P). \tag{42}$$

Direct differentiation yields

$$g'(P) = \log \frac{P(1-m)}{m(1-P)} + \alpha \frac{P - r_\beta}{P(1-P)} = F(P), \tag{43}$$

and

$$\begin{aligned}F'(P) &= \frac{1}{P(1-P)} + \alpha \frac{P^2 - 2Pr_\beta + r_\beta}{P^2(1-P)^2} \\ &= \frac{1}{P(1-P)} + \alpha \frac{(P-r_\beta)^2 + r_\beta(1-r_\beta)}{P^2(1-P)^2} > 0.\end{aligned} \tag{44}$$

Hence, g is strictly convex. The limits $F(P) \to -\infty$ as $P \downarrow 0$ and $F(P) \to +\infty$ as $P \uparrow 1$ establish a unique interior root, including when $r_\beta = 1$. The unique teacher conditionals then determine the full-support minimizer.

If $\beta = 0$, then $r_\beta = m$ and $F(m) = 0$, so $P^* = m$. If $\beta > 0$, then $r_\beta > m$ and

$$F(m) = \alpha \frac{m - r_\beta}{m(1-m)} < 0. \tag{45}$$

For $r_\beta < 1$,

$$F(r_\beta) = \log \frac{r_\beta(1-m)}{m(1-r_\beta)} > 0; \tag{46}$$

Table 3: **Support and mass control across objectives.** The final column gives the mass preferred by the forward term alone. In the combined ReTaCo objective, reverse KL also influences the equilibrium.

Objective	Reverse information	Forward support	Forward mass
Sampled reverse KL	Full KL in expectation	–	–
Renormalized forward KL	–	k in full softmax	$P \to 1$
Residual-target forward KL	–	$k+1$	$P \to r_\beta$
ReTaCo	Full KL in expectation	$k+1$	$P \to r_\beta$

when $r_\beta = 1$, the limit as $P \uparrow 1$ provides the corresponding positive upper sign. Since F is strictly increasing, the root lies in (m, r_β).

Finally,

$$\frac{\partial F}{\partial \beta} = -\alpha \frac{1-m}{P(1-P)}. \tag{47}$$

The implicit function theorem and $F'(P^*) > 0$ give

$$\frac{\mathrm{d}P^*}{\mathrm{d}\beta} = \frac{\alpha(1-m)}{P^*(1-P^*)F'(P^*)} > 0. \tag{48}$$

At both parameter endpoints, the root remains interior and the denominator finite and positive, so the one-sided derivatives are also positive. □

B.2 Forward gradient and intervention size

Proof of Proposition 3.5. Up to a teacher-only constant,

$$D_{\mathrm{KL}}(t_\beta \| C_{\mathcal{S}} p) = -\sum_{i \in \mathcal{S}} r_\beta q_i^{\mathcal{S}} \log p_i - (1 - r_\beta) \log(1-P) + \text{const}. \tag{49}$$

For $j \in \mathcal{S}$, we have $\partial P / \partial z_j = p_j(1-P)$, whereas for $j \notin \mathcal{S}$, $\partial P / \partial z_j = -P p_j$. Differentiating the two cross-entropy terms then gives

$$\frac{\partial D_{\mathrm{KL}}(t_\beta \| C_{\mathcal{S}} p)}{\partial z_j} = \begin{cases} p_j - r_\beta q_j^{\mathcal{S}}, & j \in \mathcal{S}, \\ p_j (r_\beta - P)/(1-P), & j \notin \mathcal{S}. \end{cases} \tag{50}$$

At $\beta = 0$, we have $r_\beta = m$, so the first branch approaches $-m q_j^{\mathcal{S}} = -q_j$ as $p_j \to 0$ for $j \in \mathcal{S}$. At this endpoint, both branches are zero when $p = q$. □

Target displacement. Let $\delta = r_\beta - m = \beta(1-m)$. Each selected target probability changes by $\delta q_i^{\mathcal{S}}$ relative to $C_{\mathcal{S}} q$, while the residual probability changes by $-\delta$. Therefore

$$\mathrm{TV}(t_\beta, C_{\mathcal{S}} q) = \frac{1}{2} \left(\sum_{i \in \mathcal{S}} \delta q_i^{\mathcal{S}} + \delta \right) = \delta. \tag{51}$$

The ℓ_1 change in the forward logit gradient relative to $\beta = 0$ is similarly 2δ. The absolute changes in the selected branches sum to δ, while those in the residual branches sum to $\sum_{j \notin \mathcal{S}} p_j \delta / (1-P) = \delta$. Thus, $\beta(1-m)$ directly characterizes the scale of the gradient intervention.

C Reference Implementation

C.1 Per-token computation

At each active response position, we evaluate Equation (19) as follows.

1. For the rollout token y, combine the student and teacher log-probabilities according to the straight-through estimator in Equation (17).

2. Apply log-softmax to the student logits and gather the k entries indexed by the teacher-selected IDs. Recover $\log m = \mathrm{logsumexp}_{i\in\mathcal{S}} \log q_i$ and $\log P = \mathrm{logsumexp}_{i\in\mathcal{S}} \log p_i$, then compute both residual log-probabilities with a stable $\log(1 - \exp(\cdot))$ routine.
3. Form $\log r_\beta$ and the selected target log-probabilities $\log t_i = \log r_\beta + \log q_i - \log m$.
4. Sum the selected and residual contributions for $D_{\mathrm{KL}}(t_\beta \| C_{\mathcal{S}} p)$, add the sampled reverse term, and average over active response tokens.

The rollout token, teacher-selected IDs, raw top-k log-probabilities, and set construction are detached. The student distribution remains a full-vocabulary softmax, so gradients from the aggregate tail probability also reach individual residual logits.

C.2 Numerical stability

If rounding makes selected mass indistinguishable from one, we rescale selected probabilities to sum to $1 - \varepsilon$ and assign ε to the residual. This preserves conditional proportions and normalization while perturbing aggregate mass. We use $\varepsilon = 10^{-7}$ in mixed-precision training and $\varepsilon = 0$ in float64 checks.

D Numerical Validation

D.1 Mass and Conditional-Shape Recovery

The synthetic validation jointly optimizes the selected mass and conditional distributions of a bimodal categorical student for $\beta \in \{0, 0.5, 1\}$. Table 4 compares the predicted and optimized student masses for teacher selected mass $m = 0.6802847194$.

Table 4: **Predicted and optimized selected mass.** The teacher selected mass is $m = 0.6802847194$.

Quantity	$\beta = 0$	$\beta = 0.5$	$\beta = 1$
Predicted P^*	0.6802847194	0.7642681841	0.8703486945
Optimized P	0.6802847109	0.7642681719	0.8703486945

The largest selected-conditional ℓ_1 error is 4.13×10^{-8}. Joint optimization retains the conditional tail term in Equation (13), so this test evaluates both components of the population solution.

D.2 Gradient Verification

We compare automatic differentiation with the closed-form gradients in float64 arithmetic. The calculations cover full-support distributions and distributions with a vanishing residual mass, and reproduce the equilibrium curves in Figure 3.

E Response-Length Diagnostics

E.1 Configuration

The long-run diagnostic uses a mass-preserving variant with an entropy gate, selected-token-mean forward aggregation, and a clipped PPO-style reverse objective. Table 5 gives its training configuration.

E.2 Effect of Residual-Mass Trimming

Increasing β reduces the forward target's residual mass from $1 - m$ to $(1 - \beta)(1 - m)$. When the end-of-sequence token belongs to the residual set, its probability contributes to this aggregate target. Residual trimming may therefore affect stopping behavior at prefixes where the end-of-sequence token is outside the selected support.

Table 5: Training configuration for the $\beta = 0$ response-length experiment.

Setting	Value	Setting	Value
Student	0.6B	Teacher	8B
Training examples	144,490	Optimizer steps	2,257
Learning rate	3×10^{-6}	Rollout batch	64
Samples / prompt	8	Response cap	4,096
α	1	β	0

Table 6: **Response-length behavior at $\beta = 0$.** Values are training diagnostics; response clipping denotes the fraction reaching the generation limit.

Metric	Full-run mean	Last 100	Final
Response length	477.878	470.832	482.027
Response clipping (%)	0.461	0.441	1.172
Reverse loss	1.032	0.948	0.971
Mass-preserving forward loss	1.046	0.937	1.048
Teacher top-k mass	0.999742	0.999713	0.999664
Student top-k mass	0.993973	0.9993896	0.993832

Short-run diagnostics use a 0.6B student, an 8B teacher, and a 4,096-token response budget. With the remaining settings matched, $\beta = 0$ keeps response length near 470 for 50 steps, whereas $\beta = 0.5$ exhibits response-length degeneration. Because β also changes the conditional weight r_β, this comparison does not isolate the effect of residual-mass trimming on stopping behavior.

E.3 Long-Run Behavior

The $\beta = 0$ run maintains a mean response length of approximately 478 tokens and a response clipping rate below 0.5% (Table 6).

F Experimental Details

F.1 Benchmarks and Evaluation

Table 7 lists the evaluation datasets and splits. AMC uses the 83-problem AIMO validation set adapted from AMC12 2022–2023 (Li et al., 2024).

Mathematical reasoning. We generate eight responses per problem at temperature 0.7 and top-$p = 0.95$, with no top-k truncation and a maximum of 8,192 generated tokens. Each response is scored independently using final-answer extraction and the benchmark's answer checker. For benchmark b with N_b problems, let c_i be the number of correct responses among eight samples for problem i. We report their mean correctness as avg@8

$$\text{avg@8}(b) = \frac{100}{N_b} \sum_{i=1}^{N_b} \frac{c_i}{8}. \tag{52}$$

Code generation. HumanEval+ and MBPP+ use zero-shot greedy decoding with one completion per task, temperature zero, top-$p = 1$, and an 8,192-token response limit. We score completions with the extended EvalPlus test suites.

Out-of-distribution evaluation. ARC-C, MMLU-Pro, and GPQA-Diamond are evaluated on the same epoch-1 mathematics-distilled checkpoints used for the mathematics benchmarks. OOD refers to the mathematics distillation training distribution, not the pretraining distribution. We evaluate MMLU-Pro using the official five-shot chain-of-thought protocol on the full test set. Demonstrations are drawn from the validation split of the same subject. The official evaluation configuration uses

Table 7: **Evaluation benchmarks.** HumanEval+ and MBPP+ use the full extended test suites.

Benchmark	Items	Task / split
MATH500	500	Mathematics test subset
OlympiadBench	675	English, text-only math
AMC	83	AMC12 2022–2023, integer answers
AIME24	30	2024 AIME I and II
AIME25	30	2025 AIME I and II
HumanEval+	164	EvalPlus Python tasks
MBPP+	378	EvalPlus sanitized tasks
ARC-C	1,172	Challenge test split
MMLU-Pro	12,032	Official test split
GPQA-Diamond	198	Diamond subset

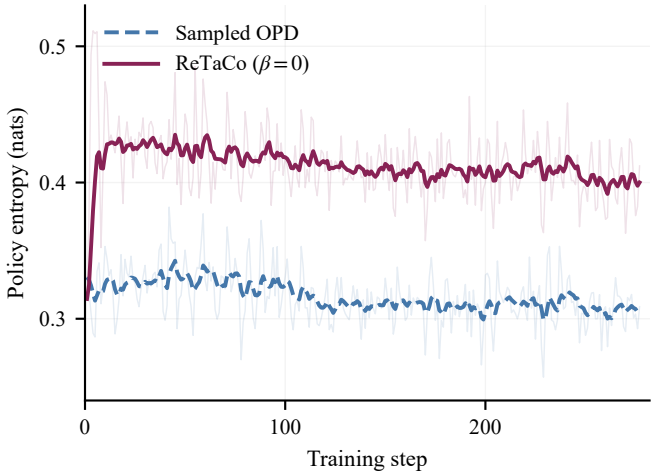


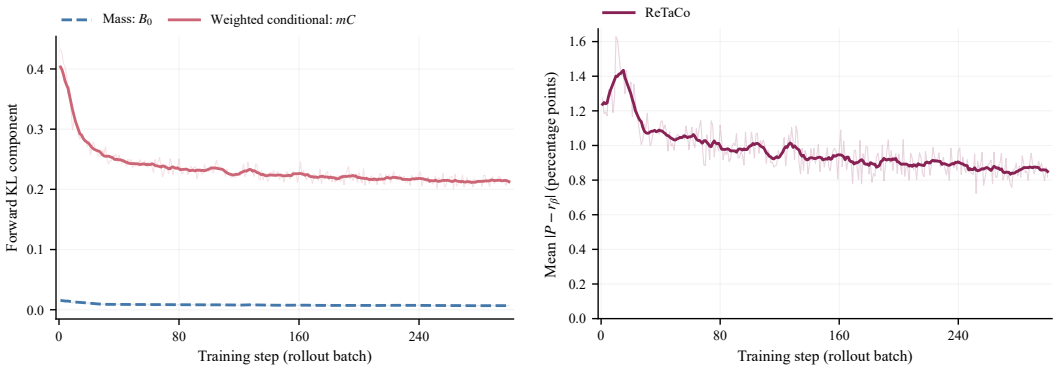
Figure 4: **Policy entropy during training.** Qwen3-1.7B student, Qwen3-30B-A3B-Instruct-2507 teacher, and $k = 16$. The curves show `actor/entropy` in nats for Sampled OPD (blue, dashed) and ReTaCo with $\beta = 0$ (wine, solid), over the common interval of steps 1–277. Solid and dashed curves use exponential smoothing with span 11; faint traces show raw values. The last 20 steps average 0.305 and 0.399 nats, respectively.

greedy decoding, a 2,048-token output limit, and the stop string `Question:`; answer extraction and accuracy follow the official evaluator (Wang et al., 2024). ARC-C and GPQA-Diamond use zero-shot generated option labels, with the same prompts and option order across methods. Invalid or missing labels count as incorrect on these two benchmarks. These evaluations measure generated-answer accuracy rather than option log-likelihood ranking.

F.2 Training-Dynamics Diagnostic

Policy entropy. Figure 4 compares Sampled OPD and ReTaCo with $\beta = 0$ over their common interval of steps 1–277. Both use the Qwen3-1.7B student, the Qwen3-30B-A3B-Instruct-2507 teacher, and $k = 16$. The recorded `actor/entropy` averages 0.399 nats for ReTaCo and 0.305 nats for Sampled OPD over the last 20 steps. ReTaCo therefore maintains higher policy entropy in this run, with a sustained separation between the two trajectories after the initial transient.

Additional loss decomposition. Figure 5 presents a Qwen3-1.7B mathematics training run with the Qwen3-30B-A3B-Instruct-2507 teacher. This seed-11 run comprises 300 consecutive rollout batches with a 7,168-token response limit, $k = 16$, $\alpha = 1$, and $\beta = 0$. It is a separate diagnostic from



(a) Forward-loss decomposition.

(b) Absolute mass error.

Figure 5: **Mass and conditional matching during training.** A single Qwen3-1.7B mathematics training run with $k = 16$, $\alpha = 1$, and $\beta = 0$. (a) Mean $B_0 = d_{\mathrm{Ber}}(m\|P)$ and $mC = mD_{\mathrm{KL}}(q^S\|p^S)$ from Equation (16). (b) Mean $|P - m|$ in percentage points. Faint lines show raw values; dark lines show centered 11-step moving averages, not confidence intervals. Steps denote rollout batches. This diagnostic uses a 7,168-token response cap and a 300-step budget, distinct from the main benchmark protocol.

the one-epoch benchmark and ablation comparisons. The curves summarize response-token statistics using micro-batch aggregation, which differs from the reduction used in the training loss.

The weighted conditional component is computed as the mean forward KL minus the mean Bernoulli KL, preserving the identity in Equation (16); it is not the product of separately averaged mass and conditional KL. Absolute mass error is the mean of per-token absolute deviations, not the absolute value of a signed mean. Initial and final values are averaged over the first and last 20 steps, respectively. For display only, an 11-step centered moving average uses the available neighborhood at each endpoint. No averaging across runs or uncertainty interval is shown.

Numerical tail projection is active at an average of 31.65% of response-token positions over the last 20 steps. These curves therefore describe the numerically stabilized implementation on evolving student-generated prefixes, rather than directly measuring the unclipped fixed-prefix population optimum. The declining conditional loss and mass error are descriptive training observations; the component ablations in Section 4.2 assess their effects on downstream accuracy.

F.3 Models and Optimization

Teacher and student models in each pair (Table 1) share compatible token vocabularies. The teacher is frozen and scores student-generated prefixes.

Training data. Mathematics training uses DAPO-Math-17k (Yu et al., 2025). Code-domain distillation uses the code subset of Eurus-2-RL-Data (Cui et al., 2025), with separate mathematics and code training runs initialized from the original student. Only problem prompts enter the distillation procedure; reference solutions and correctness rewards are not used. Each code experiment uses 8× NVIDIA H200 GPUs.

Optimization and objectives. Training settings appear in Table 8. The Qwen3 setup uses full student fine-tuning with gradient checkpointing. Sampled OPD and ReTaCo use the direct $k3+$ estimator in verl (Sheng et al., 2024), with losses averaged over valid response tokens. No task reward, reference-policy penalty, or entropy bonus is added.

Model-specific settings. The Qwen3.5 setup uses 16 H200 GPUs, with eight each for student and teacher in sampled OPD and ReTaCo; EOPD colocates its models across all 16 GPUs.

Table 8: **Mathematics training settings shared by Qwen3-1.7B, Qwen3.5-2B, and Gemma4-E2B.** $H(q)$ is full-vocabulary teacher entropy.

Setting	Value
Optimizer / learning rate	AdamW / 3×10^{-6}
AdamW betas / epsilon	$(0.9, 0.999)$ / 10^{-8}
Weight decay	0
Schedule / warmup	Cosine / 3%
Precision / gradient clipping	BF16 / 1.0
Prompt batch / mini-batch	128 / 32
Micro-batch per GPU	1
Rollouts per prompt	1
Optimization passes per batch	1
Prompt / response limits	2,048 / 8,192 tokens
Training temperature / top-p	1.0 / 1.0
Training top-k truncation	Disabled
Repetition penalty	1.0
Chat template / thinking	Native student / disabled
ReTaCo k, α, β	16, 1, 0
EOPD k, α, β	16, 1, 1
ReTaCo forward gate	None
EOPD forward gate	$H(q) > 0.8$
Loss reduction	Response-token mean
Checkpoint epoch	1

Table 9: **Response truncation in the component ablations.** Each variant trains for one epoch. Truncation denotes the fraction of responses reaching the 8,192-token generation limit.

Variant	Objective	Truncated (%)
Mass-only forward term	$R + B_0$	47.51
Conditional-only forward term	$R + mC$	48.33
No reverse term	$B_0 + mC$	49.98

EOPD baseline. EOPD uses OpenRLHF with the forward target and entropy gate in Table 8, normalized over all valid response tokens. Its reverse objective is a PPO-style surrogate with importance-ratio clipping to $[0.8, 1.2]$ and detached teacher-minus-behavior log-probability advantages. Behavior probabilities are fixed within each rollout batch.

G Additional Ablations

G.1 Component Contributions

All ablations use the Qwen3 teacher–student pair and epoch-1 checkpoints, with evaluation as in Appendix F.1.

All three component variants reach the response limit on nearly half of their completions (Table 9), indicating substantial response-length constraints.

G.2 Hyperparameter Sensitivity

Table 10 varies the forward weight and selected support. With $k = 16$ and $\beta = 0$, $\alpha = 1$ outperforms $\alpha = 0.5$ on all five tasks. Increasing α to two reduces MATH500 from 83.78% to 70.45% and AIME24 from 28.75% to 25.00%, while AIME25 increases from 21.67% to 22.50%. Stronger forward supervision thus shifts the task-level tradeoff rather than improving all benchmarks uniformly.

Table 10: **Qwen3-1.7B hyperparameter sensitivity (avg@8, %).** Every setting trains for one epoch. Defaults are $k = 16$, $\alpha = 1$, and $\beta = 0$; each group varies the indicated parameter.

Setting	MATH500	OlympiadBench	AMC	AIME24	AIME25
Forward weight					
$\alpha = 0.5$	78.45	43.78	46.84	21.25	19.17
$\alpha = 1$ (default)	83.78	49.65	52.56	28.75	21.67
$\alpha = 2$	70.45	44.15	47.59	25.00	22.50
Selected support					
$k = 8$	79.25	45.76	49.85	23.33	22.50
$k = 32$	79.35	46.17	49.10	27.50	18.33

Increasing k from eight to 32 raises AIME24 from 23.33% to 27.50%, but lowers AIME25 from 22.50% to 18.33%. The support-size tradeoff likewise varies across tasks.

H Related Work

Knowledge distillation and divergence choice. Knowledge distillation transfers teacher behavior through soft targets or generated sequences (Hinton et al., 2015; Kim & Rush, 2016). Forward KL emphasizes teacher-supported tokens, whereas reverse KL weights student-supported regions. MiniLLM adopts reverse KL (Gu et al., 2024), generalized knowledge distillation (GKD) studies divergence choice and trajectory mixtures (Agarwal et al., 2024), and other methods balance the two directions globally or token by token (Amara et al., 2022; Wu et al., 2025; Jung et al., 2025). Our analysis examines how compressing and renormalizing the teacher changes target mass even when divergence direction is fixed.

On-policy language-model distillation. On-policy distillation queries teachers at student-generated prefixes to reduce state mismatch (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025). EOPD adds forward KL at high-teacher-entropy positions to recover teacher-supported tokens that sampled reverse KL may underweight (Jin et al., 2026). ReTaCo retains EOPD's token-recovery motivation while separating selected conditional shape from aggregate mass. EOPD uses entropy gating to select supervised positions; the default ReTaCo objective applies its explicit residual target without an entropy gate.

Compressed teacher distributions. Sparse teacher targets reduce the communication and storage costs of dense teacher distributions (Shum et al., 2024; Peng et al., 2025), and tail-aware distillation separates top-k predictions from their lower-probability tail (Dasgupta et al., 2026). We use residual aggregation together with an exact diagnosis of EOPD's renormalized forward term, a single-sample estimator whose expectation equals the full-vocabulary reverse KL, and a parameterized residual target. The analysis characterizes how this target controls the joint optimum across mass preservation and deliberate trimming.

Uncertainty and reasoning diversity. Entropy-aware optimization promotes exploration at uncertain reasoning positions (Cheng et al., 2025; Wang et al., 2025). Entropy bonuses control distributional spread, whereas forward distillation supplies teacher-specific alternatives. The recovery gradient for selected teacher-supported tokens in ReTaCo follows from teacher-specific forward supervision and remains non-vanishing at $\beta = 0$, without requiring a global entropy intervention.