# TMOT: GEOMETRIC OPTIMAL TRANSPORT OVER TARGET REPRESENTATIONS FOR SPECULATIVE VERIFICATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Speculative decoding reduces the serial cost of autoregressive generation, but its gains depend on how draft tokens are verified. Token-identity and probability-only comparisons do not explicitly represent relations between distinct tokens. We formulate predictive-distribution comparison as probability-measure alignment under a ground geometry: binary token-identity cost recovers total variation, while a cost induced by target token representations replaces uniform mismatch with graded geometric mismatch. TMOT (Target-Manifold Optimal Transport) implements this view using truncated, entropy-regularized, finite-step transport. Its verifier factorizes candidate-specific probability admissibility and distribution-level geometric compatibility, with entropy and sharpness adapting the allowable discrepancy. The resulting module is training-free and can be integrated into a speculative decoding pipeline without changing model parameters. Code-generation and mathematical-reasoning experiments yield up to $1.90\times$ end-to-end speedup, with task-dependent accuracy recovery and a small measured verification-time share. These results support a quality–efficiency trade-off for geometric relaxation; TMOT is a lossy verifier and does not preserve the exact target sampling distribution.

## 1 INTRODUCTION

Autoregressive language models generate tokens sequentially, making the cost of repeated target-model evaluation a central inference bottleneck. Speculative decoding amortizes this cost by using a smaller draft model to propose several tokens and scoring them with one target-model forward pass. With an appropriate acceptance and correction rule, speculative sampling preserves the target distribution (Leviathan et al., 2023). Improvements to drafting, including multiple decoding heads and feature prediction, increase the usefulness of these proposals (Cai et al., 2024; Li et al., 2024).

Verification determines how much of each proposal can be retained. Greedy verification compares the proposal with the target argmax; distribution-preserving speculative sampling instead uses probability ratios and correction sampling. These are distinct rules, and neither should be described as a universal exact-token-match requirement. Both motivate examining how draft and target predictive distributions are compared. Token-identity and probability-only discrepancies do not explicitly distinguish nearby from distant representations of different tokens.

Relaxed verification already explores alternative quality–efficiency trade-offs. FSD uses divergences between predictive distributions (Holsman et al., 2025), and FLy combines an entropy gate with deferred verification (Li et al., 2026). Optimal transport (OT) also already appears in speculative decoding: SpecTr uses it for draft selection while retaining the target distribution (Sun et al., 2023). These developments motivate a specific geometric question:

> *Can speculative verification be formulated as probability-measure alignment over the target model's own representation geometry?*

We develop this view through TMOT (Target-Manifold Optimal Transport). An identity ground cost charges one unit for every pair of different tokens; its ideal transport discrepancy equals total variation. Replacing that cost with normalized target-embedding cosine cost gives graded geometric



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.

mismatch. TMOT treats the target model's representation space as the ground geometry over which draft and target predictive measures are compared. Here, "Target Manifold" names a discrete geometry induced by static input embeddings, without assuming a continuous manifold or complete contextual semantics.

Geometry alone does not determine whether a proposed token is admissible. TMOT therefore separates a candidate-specific probability gate from a distribution-level geometric state gate. The former requires local target support; the latter determines whether the predictive state permits relaxation. Entropy and sharpness calibrate the allowable geometric discrepancy. This factorization makes the geometric score's candidate independence explicit. Our contributions are:

- **A geometric view of speculative verification.** We formulate predictive-measure alignment under a ground cost and relate binary token-identity geometry to total variation.
- **Target-induced transport geometry.** We replace binary mismatch with target-representation cost, derive an ideal-OT comparison bound, and distinguish it from the practical transport score.
- **Factorized geometric verification.** We express candidate admissibility and predictive-state compatibility separately, establishing support constraints and monotone uncertainty adaptation of the acceptance threshold.

Experiments on HumanEval, MBPP, GSM8K, and four MGSM languages assess end-to-end performance, verification overhead, component contributions, and hyperparameter sensitivity. They demonstrate task-dependent quality–efficiency trade-offs, including accuracy losses.

## 2 Geometric Speculative Verification

### 2.1 Problem Setup

At verification position $i$ in a block of $N$ draft tokens ($N = 8$ for TMOT), let $P_D^{(i)}$ and $P_T^{(i)}$ denote draft and target conditional distributions on a shared token vocabulary $\mathcal{V}$. They are evaluated along the proposed prefix, which becomes valid only if all preceding tokens are accepted. Let $x_i^d$ be the proposed token and $x_i^t = \arg\max_x P_T^{(i)}(x)$ the target greedy token, using a consistent tie rule. TMOT accepts exact matches directly and applies relaxed verification otherwise. Token identities must be aligned across the two models; no cross-tokenizer mapping is introduced here.

### 2.2 From Discrete Verification to Geometric Verification

For probability vectors $\mu_d, \mu_t$ on a finite token support $U$, let

$$\Pi(\mu_d, \mu_t) = \{T \geq 0 : T\mathbf{1} = \mu_d,\ T^\top \mathbf{1} = \mu_t\}, \qquad W_C(\mu_d, \mu_t) = \min_{T \in \Pi(\mu_d, \mu_t)} \langle T, C \rangle_F. \quad (1)$$

The ground cost $C(x, y)$ specifies the penalty for aligning probability mass on $x$ with mass on $y$ (Peyré & Cuturi, 2019). This is a discrepancy view of predictive distributions, not an equivalence between an entire speculative decoding algorithm and OT.

**Proposition 1** (Binary token-identity geometry)**.** *For* $C_{\text{disc}}(x, y) = \mathbb{1}[x \neq y]$,

$$W_{C_{\text{disc}}}(\mu_d, \mu_t) = 1 - \sum_{x \in U} \min\{\mu_d(x), \mu_t(x)\} = \text{TV}(\mu_d, \mu_t), \quad (2)$$

*where* $\text{TV}(\mu_d, \mu_t) = \frac{1}{2} \sum_x |\mu_d(x) - \mu_t(x)|$.

Thus all non-identical tokens receive the same unit transportation cost. This coarse geometry records identity but no graded relation among distinct tokens. Appendix A gives proofs of all four propositions.

### 2.3 Target-Induced Representation Geometry

Let $\mathbf{e}(x)$ be a nonzero static target input embedding, and set $\bar{\mathbf{e}}(x) = \mathbf{e}(x)/\|\mathbf{e}(x)\|_2$. Define

$$C_{\text{TMOT}}(x, y) = \frac{1}{2}\left(1 - \langle \bar{\mathbf{e}}(x), \bar{\mathbf{e}}(y) \rangle\right). \quad (3)$$

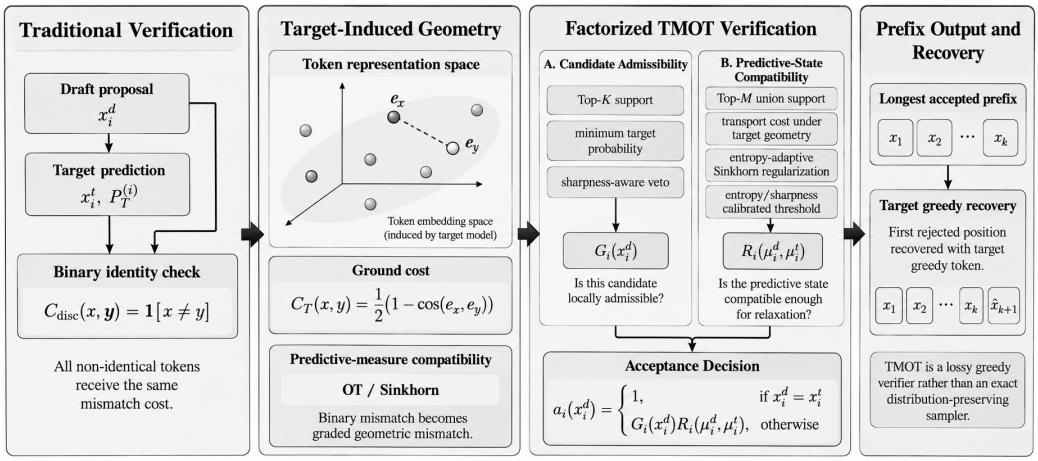


Figure 1: Geometric formulation and factorized TMOT verification. Binary token-identity cost is replaced by target-induced ground geometry. Non-exact proposals require candidate admissibility and predictive-state compatibility; entropy and sharpness calibrate the geometric threshold. Prefix acceptance is followed by target greedy recovery at the first rejection.

The cost uses target representations for both measures and does not require draft embeddings to share their coordinate system.

**Proposition 2** (Graded geometric mismatch)**.** *For unit-normalized representations,* $0 \leq C_{\text{TMOT}}(x, y) \leq 1$ *and* $C_{\text{TMOT}}(x, x) = 0$*. Consequently,* $C_{\text{TMOT}} \leq C_{\text{disc}}$ *pointwise and, for ideal exact OT on the same probability measures,*

$$W_{C_{\text{TMOT}}}(\mu_d, \mu_t) \leq W_{C_{\text{disc}}}(\mu_d, \mu_t) = \text{TV}(\mu_d, \mu_t). \tag{4}$$

This cost need not be a metric: distinct tokens can share normalized representations, and the triangle inequality can fail. OT accounts for redistribution of probability mass across entire measures, beyond the cosine similarity of a single token pair. The comparison bound concerns geometry and does not establish semantic correctness or downstream error control.

### 2.4 Factorized Geometric Verification

Write $G_i(x_i^d)$ for candidate admissibility and $R_i(\mu_i^d, \mu_i^t)$ for predictive-state geometric compatibility, with target uncertainty understood as part of the state. The verification rule is

$$a_i(x_i^d) = \begin{cases} 1, & x_i^d = x_i^t, \\ G_i(x_i^d) R_i(\mu_i^d, \mu_i^t), & x_i^d \neq x_i^t. \end{cases} \tag{5}$$

Fixing the predictive distributions fixes the OT score; candidate identity enters the exact-match branch and $G_i$. Figure 1 summarizes these two levels of verification.

### 2.5 Practical Target-Geometry Transport

**Truncated predictive measures.** Let $\mathcal{S}_i^d = \text{TopM}(P_D^{(i)})$, $\mathcal{S}_i^t = \text{TopM}(P_T^{(i)})$, and $\mathcal{U}_i = \mathcal{S}_i^d \cup \mathcal{S}_i^t = \{x_{i,1}, \ldots, x_{i,n_i}\}$, where $n_i \leq 2M$ and the default is $M = 32$. We first form the union of the draft and target Top-$M$ token indices, retrieve both models' original probability masses on this shared support, and renormalize each measure independently:

$$\tilde{p}_{i,a}^q = \frac{P_q^{(i)}(x_{i,a})}{\sum_{x \in \mathcal{U}_i} P_q^{(i)}(x)}, \qquad \mu_i^q = \sum_{a=1}^{n_i} \tilde{p}_{i,a}^q \delta_{x_{i,a}}, \quad q \in \{D, T\}. \tag{6}$$

The vector forms are denoted $\boldsymbol{\mu}_i^d$, $\boldsymbol{\mu}_i^t$, and $C_{i,ab} = C_{\text{TMOT}}(x_{i,a}, x_{i,b})$.

**Entropy-adaptive regularization (AESA).** For $J = M$ or $K$, define the raw truncated entropy and its scaled value by

$$\mathcal{H}^{\mathrm{raw}}_{i,J} = -\sum_{x \in \mathrm{TopJ}(P_T^{(i)})} P_T^{(i)}(x) \log P_T^{(i)}(x), \qquad \bar{\mathcal{H}}_{i,J} = \frac{\mathcal{H}^{\mathrm{raw}}_{i,J}}{\log J + 10^{-9}}. \tag{7}$$

These use original target probabilities, without renormalizing the Top-$J$ mass; entropy normalization retains a $10^{-9}$ numerical stabilizer. For $J \geq 3$, $0 \leq \bar{\mathcal{H}}_{i,J} \leq 1$ (Appendix B); they are uncertainty signals, not the entropy of the normalized transport measure. The regularization is

$$\epsilon_i^{\mathrm{dyn}} = \frac{\epsilon_{\mathrm{base}}}{2}\left(1 + \alpha \bar{\mathcal{H}}_{i,M}\right), \qquad \epsilon_{\mathrm{base}} = 0.05, \quad \alpha = 0.5. \tag{8}$$

Higher local uncertainty increases regularization, encouraging a more diffuse plan; it does not imply that the resulting transport cost decreases.

**Finite-step Sinkhorn score.** Entropic OT minimizes $\langle T, C_i \rangle_F + \epsilon_i^{\mathrm{dyn}} \sum_{ab} T_{ab}(\log T_{ab} - 1)$ over the marginal constraints, with $0 \log 0 = 0$ (Cuturi, 2013). TMOT approximates the associated scaling using $\mathbf{K}_i = \exp(-\mathbf{C}_i/\epsilon_i^{\mathrm{dyn}})$ and $\mathbf{u}_i^{(0)} = \mathbf{v}_i^{(0)} = \mathbf{1}/n_i$:

$$\mathbf{u}_i^{(r+1)} = \frac{\boldsymbol{\mu}_i^d}{\mathbf{K}_i \mathbf{v}_i^{(r)} + \delta}, \qquad \mathbf{v}_i^{(r+1)} = \frac{\boldsymbol{\mu}_i^t}{\mathbf{K}_i^\top \mathbf{u}_i^{(r+1)} + \delta}. \tag{9}$$

Divisions are elementwise, $\delta = 10^{-12}$, and $L = 5$ iterations are used. With $\mathbf{T}_i^{(L)} = \mathrm{diag}(\mathbf{u}_i^{(L)})\mathbf{K}_i \,\mathrm{diag}(\mathbf{v}_i^{(L)})$, the score is

$$D_i \equiv D^{(i)}_{\mathrm{TMOT}} = \langle \mathbf{T}_i^{(L)}, \mathbf{C}_i \rangle_F. \tag{10}$$

The reported score contains the transport-cost term, not the entropy term of the optimization objective.

**From ideal geometric transport to practical TMOT.** Propositions 1–2 concern exact finite-support OT with feasible marginal constraints. Practical TMOT uses a Top-$M$ union, truncated and renormalized mass, normalized target embeddings, entropy regularization, fixed Sinkhorn iterations, and early candidate veto. Truncation changes the measures, regularization changes the optimizer, and finite iterations with stabilizers need not satisfy the marginals exactly. Thus $D_i$ is a geometric compatibility signal, not an exact Wasserstein distance. In particular, equation 4 is not an upper bound on $D_i$ or on discrepancy between the original full-vocabulary distributions.

### 2.6 Support-Constrained Candidate Admissibility

Let $\mathcal{S}_{i,K}^t = \mathrm{TopK}(P_T^{(i)})$, with default $K = 3$, and define

$$p^t_{i,\max} = \max_x P_T^{(i)}(x), \qquad \mathcal{S}_i = \frac{p^t_{i,\max}}{\sum_{x \in \mathcal{S}^t_{i,K}} P_T^{(i)}(x)} \in [1/K, 1]. \tag{11}$$

The Fatal Veto gate is equivalently written as the following conjunction:

$$\begin{aligned} G_i(x) = \mathbb{1}[x \in \mathcal{S}^t_{i,K}]\, \mathbb{1}[P_T^{(i)}(x) \geq \eta] \\ \cdot\, \mathbb{1}[\mathcal{S}_i \leq \tau_{\mathrm{sharp}} \ \vee \ P_T^{(i)}(x) \geq \rho p^t_{i,\max}], \end{aligned} \tag{12}$$

where $\eta = 0.02$, $\rho = 0.5$, and $\tau_{\mathrm{sharp}} = 0.8$. A non-exact candidate with $G_i = 0$ is rejected before constructing the transport support.

**Proposition 3** (Support-constrained relaxed acceptance). *If $x_i^d \neq x_i^t$ and $a_i(x_i^d) = 1$, then $x_i^d \in \mathrm{TopK}(P_T^{(i)})$ and $P_T^{(i)}(x_i^d) \geq \eta$. If additionally $\mathcal{S}_i > \tau_{\mathrm{sharp}}$, then $P_T^{(i)}(x_i^d) \geq \rho p^t_{i,\max}$.*

These are deterministic constraints inherited from the gate. They provide no calibrated probability of correctness, bound on sequence-level error, or semantic safety guarantee.

### 2.7 UNCERTAINTY-CONSISTENT ACCEPTANCE

The entropy penalty and sharpness-dependent tolerance are

$$\Phi_i = \exp[\gamma \max(0, \theta - \bar{\mathcal{H}}_{i,K})], \qquad \tau_i^{\mathrm{dyn}} = \tau_{\mathrm{ot}}(1 - \mathcal{S}_i^2) + \varepsilon, \tag{13}$$

with $\theta = 0.5$, $\gamma = 4.0$, $\tau_{\mathrm{ot}} = 0.15$, and $\varepsilon = 10^{-9}$. The predictive-state compatibility gate, with default $\kappa = 1$, is

$$R_i = \mathbb{1}\left[\exp\left(-\frac{D_i \Phi_i}{\tau_i^{\mathrm{dyn}}/2}\right) > e^{-\kappa}\right] = \mathbb{1}\left[D_i \Phi_i < \frac{\kappa \tau_i^{\mathrm{dyn}}}{2}\right]. \tag{14}$$

For $H = \bar{\mathcal{H}}_{i,K}$ and $S = \mathcal{S}_i$, this is exactly $D_i < B(H, S)$, where

$$B(H, S) = \frac{\kappa}{2}[\tau_{\mathrm{ot}}(1 - S^2) + \varepsilon] \exp[-\gamma \max(0, \theta - H)]. \tag{15}$$

**Proposition 4** (Monotone uncertainty adaptation)**.** *For $\kappa > 0$, $\tau_{\mathrm{ot}} \geq 0$, $\gamma \geq 0$, $\varepsilon > 0$, and $S \in [0, 1]$, the allowable discrepancy $B$ is nondecreasing in $H$ and nonincreasing in $S$. Where differentiable,*

$$\frac{\partial B}{\partial H} \geq 0, \qquad \frac{\partial B}{\partial S} \leq 0. \tag{16}$$

*At $H = \theta$, the same monotonicity holds with one-sided derivatives.*

The threshold increases with entropy below $\theta$, is constant above it, and decreases with sharpness. This coordinatewise property does not imply monotone end-to-end acceptance: changing a distribution also changes $D_i$, AESA, and candidate admissibility.

### 2.8 UNIFIED ALGORITHM AND COMPLEXITY

Only the longest valid prefix is retained:

$$L_{\mathrm{acc}} = \max\{\ell \in \{0, \ldots, N\} : a_1 = \cdots = a_\ell = 1\}. \tag{17}$$

The implementation evaluates position-wise gates in parallel and applies a cumulative prefix mask; Algorithm 1 presents the equivalent sequential semantics for clarity. Specifically, `torch.cummin(final_accepted, dim=1).values` enforces prefix consistency. At the first rejection, TMOT's `_verify_tokens` recovers the position using the target model's greedy prediction. This rule performs no rejection sampling; TMOT is evaluated as a relaxed greedy verifier and does not preserve the exact target sampling distribution.

Excluding model forward passes and vocabulary-level support selection, the geometric work per eligible position is $O(n_i^2 d + L n_i^2)$, with $n_i \leq 2M$, and uses $O(n_i^2 + n_i d)$ working storage. Exact-match and veto branches skip this work. This bound describes the geometric stage, not total decoding complexity. Table 2 reports measured overhead.

## 3 EXPERIMENTS

We evaluate end-to-end efficiency and task performance (RQ1), verification overhead (RQ2), component contributions (RQ3), and hyperparameter sensitivity (RQ4).

### 3.1 EXPERIMENTAL SETUP

**Models and tasks.** We evaluate three draft–target pairs: Llama-3.2-3B → Llama-3.1-13B (Pair A), Llama-3.2-3B → Llama-3.2-8B (Pair B), and DeepSeek-R1-Distill-Qwen-1.5B → DeepSeek-R1-Distill-Qwen-7B (Pair C; abbreviated DSR1-Qwen). Exact checkpoint IDs are listed in Appendix B. Tasks are HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), GSM8K (Cobbe et al., 2021), and MGSM in English, Chinese, German, and French (Shi et al., 2022).

**Algorithm 1** TMOT verification: sequential semantics of vectorized prefix masking

**Require:** Draft tokens $x_1^d, \ldots, x_N^d$, draft distributions $P_D$, target embeddings $\mathbf{E}$

Obtain $P_T$ with one batched target forward pass along the draft prefix
$L_{\text{acc}} \leftarrow 0$
**for** $i = 1$ to $N$ **do**
  $x_i^t \leftarrow \arg\max_x P_T^{(i)}(x)$
  **if** $x_i^d \neq x_i^t$ **then**
    Compute $\mathcal{S}_i$ and $G_i(x_i^d)$ using equation 12
    **if** $G_i(x_i^d) = 0$ **then**
      **break** ▷ Candidate veto
    **end if**
    Gather and renormalize $\mu_i^d, \mu_i^t$ on the Top-$M$ union; construct $\mathbf{C}_i$
    Compute $\epsilon_i^{\text{dyn}}$ and perform $L$ Sinkhorn iterations to obtain $D_i$
    Compute $\Phi_i$ and $\tau_i^{\text{dyn}}$
    **if** $D_i \Phi_i \geq \kappa \tau_i^{\text{dyn}}/2$ **then**
      **break** ▷ State rejection
    **end if**
  **end if**
  $L_{\text{acc}} \leftarrow i$
**end for**
Retain $x_1^d, \ldots, x_{L_{\text{acc}}}^d$
**if** $L_{\text{acc}} < N$ **then**
  Recover the first rejected position with $\arg\max_x P_T^{(L_{\text{acc}}+1)}(x)$
**else**
  Append one target-generated bonus token after the fully accepted block
**end if**

**Environment and decoding.** Experiments use four NVIDIA Tesla V100S GPUs with 32 GB each, dual Intel Xeon Gold 6248R CPUs, Ubuntu 22.04.3, PyTorch 2.4.0, and Transformers 4.44.0. Inference uses FP16, tensor parallelism of four, no data parallelism, and batch size one. We use greedy decoding with temperature 0.0 and seed 42. TMOT and the vLLM speculative baseline both use a draft block length of 8. Generation is limited to 512 new tokens with a maximum context length of 4096. We compare target-only autoregressive decoding (Baseline), vLLM, FSD, PEARL, FLy, and TMOT. We use vLLM 0.5.5 built-in speculative decoding as an implementation baseline under the same greedy regime and seed.

**Metrics and timing.** Tokens/s is workload-level generation throughput, and speedup is its ratio to target-only throughput. Mean Advanced Tokens (MAT) denotes the average number of tokens advanced per speculative round, including accepted draft tokens and one target-generated bonus or correction token. Thus $\text{MAT} = R^{-1} \sum_{r=1}^{R} (A_r + 1)$, where $A_r$ is the number of accepted draft tokens in round $r$ and $R$ is the number of rounds. ACC is pass@1 for code generation and exact-match accuracy for mathematical reasoning. Recovery is $100\,\text{ACC}_{\text{method}}/\text{ACC}_{\text{baseline}}$; displayed ratios use the displayed rounded measurements. Timing uses warm-up and CUDA synchronization; Table 2 separates draft, target, and verification time.

### 3.2 RQ1: End-to-End Efficiency and Task Performance

Table 1 reports up to $1.90\times$ speedup (51.03 tokens/s, MGSM-EN, Llama-3.2-3B $\rightarrow$ Llama-3.1-13B), with 62.00 ACC versus 62.80 for target-only decoding, or 98.73% recovery. Across its 21 TMOT model–task entries, speedup ranges from $1.13\times$ to $1.90\times$, and recovery from 89.27% to 100.00%. The lowest recovery occurs on Llama-3.2-8B/MGSM-ZH: ACC falls from 70.80 to 63.20. DSR1-Qwen-7B/MGSM-ZH also yields 95.52% recovery. Recovery therefore varies substantially across tasks.

The comparisons reflect task-dependent trade-offs. For example, DSR1-Qwen-7B/HumanEval has TMOT throughput 30.10 versus FLy's 30.17 tokens/s, but higher ACC (64.02 versus 60.98). On

Table 1: End-to-end results for the three target models, paired with their drafters in Section 3.1. Rate is tokens/s (speedup); ACC is in percent with relative recovery in parentheses. MAT denotes Mean Advanced Tokens. Ratios use displayed rounded measurements. Bold identifies TMOT. DSR1-Qwen abbreviates DeepSeek-R1-Distill-Qwen.

| | | Llama-3.1-13B | | | Llama-3.2-8B | | | DSR1-Qwen-7B | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Method** | **Rate (×)** | **MAT** | **ACC (Rec)** | **Rate (×)** | **MAT** | **ACC (Rec)** | **Rate (×)** | **MAT** | **ACC (Rec)** |
| HumanEval | Baseline | 23.67 (1.00×) | - | 50.00 (—) | 39.72 (1.00×) | - | 71.34 (—) | 23.76 (1.00×) | - | 64.02 (—) |
| | vLLM | 40.65 (1.72×) | 6.35 | 46.95 (93.90%) | 51.75 (1.30×) | 6.21 | 70.73 (99.14%) | 28.11 (1.18×) | 4.35 | 64.02 (100.00%) |
| | FSD | 37.71 (1.59×) | 6.42 | 46.95 (93.90%) | 52.31 (1.32×) | 6.20 | 69.36 (97.22%) | 27.89 (1.17×) | 4.53 | 64.02 (100.00%) |
| | PEARL | 37.12 (1.57×) | 6.57 | 46.95 (93.90%) | 58.17 (1.46×) | 6.24 | 64.02 (89.74%) | 25.96 (1.09×) | 4.25 | 60.59 (94.64%) |
| | FLy | 40.94 (1.73×) | 6.77 | 48.17 (96.34%) | 52.72 (1.33×) | 6.30 | 68.59 (96.15%) | 30.17 (1.27×) | 4.67 | 60.98 (95.25%) |
| | **Ours** | **42.73 (1.81×)** | **6.79** | **48.78 (97.56%)** | **58.23 (1.47×)** | **6.57** | **69.51 (97.43%)** | **30.10 (1.27×)** | **4.61** | **64.02 (100.00%)** |
| GSM8K | Baseline | 25.71 (1.00×) | - | 52.39 (—) | 40.74 (1.00×) | - | 77.48 (—) | 27.81 (1.00×) | - | 89.99 (—) |
| | vLLM | 40.82 (1.59×) | 6.26 | 52.16 (99.56%) | 40.82 (1.00×) | 5.52 | 77.33 (99.81%) | 34.76 (1.25×) | 5.72 | 89.11 (99.02%) |
| | FSD | 41.75 (1.62×) | 6.25 | 52.16 (99.56%) | 48.53 (1.19×) | 5.77 | 76.24 (98.40%) | 34.62 (1.24×) | 5.72 | 87.74 (97.50%) |
| | PEARL | 38.34 (1.49×) | 6.09 | 48.06 (91.74%) | 47.14 (1.16×) | 5.54 | 70.96 (91.58%) | 35.58 (1.28×) | 5.69 | 83.85 (93.18%) |
| | FLy | 39.45 (1.53×) | 6.24 | 49.41 (94.31%) | 46.59 (1.14×) | 5.72 | 72.18 (93.16%) | 35.18 (1.27×) | 5.85 | 86.35 (95.96%) |
| | **Ours** | **42.38 (1.65×)** | **6.56** | **52.24 (99.71%)** | **53.09 (1.30×)** | **6.19** | **76.50 (98.74%)** | **36.67 (1.32×)** | **6.00** | **88.93 (98.82%)** |
| MBPP | Baseline | 22.33 (1.00×) | - | 40.20 (—) | 36.11 (1.00×) | - | 60.20 (—) | 26.59 (1.00×) | - | 34.40 (—) |
| | vLLM | 35.74 (1.60×) | 6.35 | 40.20 (100.00%) | 49.17 (1.36×) | 5.56 | 60.20 (100.00%) | 29.90 (1.12×) | 4.74 | 33.80 (98.26%) |
| | FSD | 34.61 (1.55×) | 6.42 | 40.00 (99.50%) | 47.57 (1.32×) | 5.45 | 57.20 (95.02%) | 27.93 (1.05×) | 4.74 | 32.60 (94.77%) |
| | PEARL | 37.72 (1.69×) | 6.02 | 38.40 (95.52%) | 47.22 (1.31×) | 5.71 | 54.00 (89.70%) | 28.95 (1.09×) | 4.89 | 31.40 (91.28%) |
| | FLy | 37.05 (1.66×) | 6.70 | 39.60 (98.51%) | 50.16 (1.39×) | 5.77 | 58.60 (97.34%) | 29.71 (1.12×) | 4.92 | 32.60 (94.77%) |
| | **Ours** | **37.76 (1.69×)** | **6.48** | **40.20 (100.00%)** | **53.16 (1.47×)** | **5.78** | **60.20 (100.00%)** | **31.19 (1.17×)** | **4.94** | **34.40 (100.00%)** |
| MGSM-ZH | Baseline | 25.73 (1.00×) | - | 34.80 (—) | 42.86 (1.00×) | - | 70.80 (—) | 28.69 (1.00×) | - | 80.40 (—) |
| | vLLM | 36.00 (1.40×) | 5.72 | 34.40 (98.85%) | 52.14 (1.22×) | 5.42 | 70.40 (99.44%) | 36.22 (1.26×) | 5.76 | 75.60 (94.03%) |
| | FSD | 35.03 (1.36×) | 5.62 | 34.40 (98.85%) | 48.18 (1.12×) | 5.43 | 63.20 (89.27%) | 35.80 (1.25×) | 5.76 | 75.60 (94.03%) |
| | PEARL | 35.24 (1.37×) | 5.23 | 32.80 (94.25%) | 50.03 (1.17×) | 5.29 | 62.80 (88.70%) | 37.18 (1.30×) | 5.88 | 69.20 (86.07%) |
| | FLy | 36.68 (1.43×) | 5.72 | 33.60 (96.55%) | 47.04 (1.10×) | 5.47 | 64.80 (91.53%) | 34.39 (1.20×) | 5.95 | 74.20 (92.29%) |
| | **Ours** | **39.13 (1.52×)** | **5.82** | **34.80 (100.00%)** | **50.14 (1.17×)** | **5.93** | **63.20 (89.27%)** | **38.80 (1.35×)** | **6.03** | **76.80 (95.52%)** |
| MGSM-EN | Baseline | 26.88 (1.00×) | - | 62.80 (—) | 43.07 (1.00×) | - | 82.80 (—) | 29.23 (1.00×) | - | 86.00 (—) |
| | vLLM | 46.17 (1.72×) | 6.72 | 62.00 (98.73%) | 56.86 (1.32×) | 6.06 | 81.20 (98.07%) | 37.56 (1.28×) | 5.94 | 85.60 (99.53%) |
| | FSD | 47.66 (1.77×) | 6.58 | 62.00 (98.73%) | 57.77 (1.34×) | 6.04 | 80.00 (96.62%) | 38.75 (1.33×) | 5.94 | 82.80 (96.28%) |
| | PEARL | 42.46 (1.58×) | 6.52 | 56.80 (90.45%) | 55.88 (1.30×) | 6.17 | 79.50 (96.01%) | 39.15 (1.34×) | 6.04 | 79.20 (92.09%) |
| | FLy | 47.12 (1.75×) | 6.63 | 57.60 (91.72%) | 60.20 (1.40×) | 6.35 | 80.00 (96.62%) | 37.25 (1.27×) | 6.05 | 82.80 (96.28%) |
| | **Ours** | **51.03 (1.90×)** | **6.75** | **62.00 (98.73%)** | **59.79 (1.39×)** | **6.75** | **81.20 (98.07%)** | **39.14 (1.34×)** | **6.16** | **84.00 (97.67%)** |
| MGSM-DE | Baseline | 26.78 (1.00×) | - | 47.60 (—) | 45.23 (1.00×) | - | 68.80 (—) | 28.71 (1.00×) | - | 45.20 (—) |
| | vLLM | 43.38 (1.62×) | 6.37 | 47.60 (100.00%) | 43.38 (0.96×) | 5.64 | 68.80 (100.00%) | 32.08 (1.12×) | 5.02 | 45.20 (100.00%) |
| | FSD | 44.77 (1.67×) | 6.49 | 47.60 (100.00%) | 52.37 (1.16×) | 5.62 | 64.80 (94.19%) | 29.05 (1.01×) | 5.02 | 45.20 (100.00%) |
| | PEARL | 37.31 (1.39×) | 5.49 | 43.40 (91.18%) | 49.06 (1.08×) | 5.21 | 61.00 (88.66%) | 33.05 (1.15×) | 5.21 | 43.20 (95.58%) |
| | FLy | 37.94 (1.42×) | 6.63 | 46.80 (98.32%) | 50.27 (1.11×) | 5.47 | 64.80 (94.19%) | 33.18 (1.16×) | 5.13 | 44.80 (99.12%) |
| | **Ours** | **44.71 (1.67×)** | **6.51** | **47.60 (100.00%)** | **51.85 (1.15×)** | **6.03** | **68.80 (100.00%)** | **32.37 (1.13×)** | **5.15** | **44.80 (99.12%)** |
| MGSM-FR | Baseline | 27.69 (1.00×) | - | 46.40 (—) | 43.56 (1.00×) | - | 68.80 (—) | 28.50 (1.00×) | - | 65.60 (—) |
| | vLLM | 43.41 (1.57×) | 6.33 | 46.40 (100.00%) | 55.34 (1.27×) | 5.81 | 68.80 (100.00%) | 32.81 (1.15×) | 5.15 | 64.80 (98.78%) |
| | FSD | 45.20 (1.63×) | 6.49 | 46.40 (100.00%) | 56.16 (1.29×) | 5.83 | 61.00 (88.66%) | 33.35 (1.17×) | 5.15 | 65.20 (99.39%) |
| | PEARL | 39.19 (1.42×) | 5.84 | 42.00 (90.52%) | 51.71 (1.19×) | 5.50 | 60.00 (87.21%) | 34.06 (1.20×) | 5.33 | 64.00 (97.56%) |
| | FLy | 41.54 (1.50×) | 5.96 | 43.20 (93.10%) | 54.06 (1.24×) | 5.96 | 62.80 (91.28%) | 33.96 (1.19×) | 5.28 | 65.60 (100.00%) |
| | **Ours** | **46.12 (1.67×)** | **6.70** | **45.60 (98.28%)** | **57.45 (1.32×)** | **6.27** | **68.80 (100.00%)** | **33.49 (1.18×)** | **5.29** | **65.20 (99.39%)** |

Llama-3.2-8B/MGSM-ZH, the reported vLLM row is both faster and more accurate than TMOT. These individual operating points characterize trade-offs without establishing Pareto dominance.

### 3.3 RQ2: Verification Overhead

Table 2 separates draft, target, and verification times. Verification accounts for 1.61–2.80% of time in the nine task-specific rows; the three model-pair aggregates are 2.22%, 2.51%, and 1.62%. For DSR1-Qwen, verification totals 172.31 s out of 10646.92 s. These measurements indicate a small verification share on the reported workloads. The bounded support and early gate limit geometric work, but the table does not isolate the speed contribution of each implementation choice or establish overhead on other hardware.

### 3.4 RQ3: Component Contributions

Table 3 evaluates three interventions. **w/o Veto** disables the fatal-hallucination mask while retaining sharpness adaptation. **w/o AESA** replaces entropy-adaptive Sinkhorn regularization with fixed $\epsilon_{\text{base}}/2$, retaining entropy penalty and sharpness adaptation. **w/o Penalty** sets $\Phi_i = 1$, retaining AESA and sharpness adaptation.

Table 2: Phase timings. Verify (%) is verification time divided by total time. Total rows aggregate the three displayed tasks for each model pair.

| Draft → Target | Dataset | Draft (s) | Target (s) | Verify (s) | Total (s) | Verify (%) |
|---|---|---|---|---|---|---|
| Llama-3.2-3B → Llama-3.1-13B | HumanEval | 246.75 | 195.84 | 10.30 | 452.89 | 2.27 |
| | GSM8K | 2656.60 | 2117.14 | 108.60 | 4882.34 | 2.22 |
| | MGSM-EN | 656.10 | 551.87 | 26.80 | 1234.77 | 2.17 |
| | Total | 3559.45 | 2864.85 | 145.70 | 6570.00 | 2.22 |
| Llama-3.2-3B → Llama-3.2-8B | HumanEval | 153.17 | 91.58 | 5.95 | 250.70 | 2.37 |
| | GSM8K | 1349.39 | 806.48 | 53.13 | 2209.00 | 2.41 |
| | MGSM-EN | 628.52 | 311.65 | 27.07 | 967.24 | 2.80 |
| | Total | 2131.08 | 1209.71 | 86.15 | 3426.94 | 2.51 |
| DSR1-Qwen-1.5B → DSR1-Qwen-7B | HumanEval | 379.69 | 180.17 | 10.43 | 570.29 | 1.83 |
| | GSM8K | 5845.27 | 2757.79 | 140.41 | 8743.47 | 1.61 |
| | MGSM-EN | 896.91 | 414.78 | 21.47 | 1333.16 | 1.61 |
| | Total | 7121.87 | 3352.74 | 172.31 | 10646.92 | 1.62 |

Table 3: Component ablations defined in Section 3.4. Rate is tokens/s (speedup); ACC includes relative recovery. All three variants retain sharpness adaptation and OT. Ours denotes full TMOT.

| | | HumanEval | | | GSM8K | | | MGSM-EN | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Method | Rate (×) | MAT | ACC (Rec) | Rate (×) | MAT | ACC (Rec) | Rate (×) | MAT | ACC (Rec) |
| Llama-3.1-13B | Baseline | 23.67 (1.00×) | - | 50.00 (—) | 25.71 (1.00×) | - | 52.39 (—) | 26.88 (1.00×) | - | 62.80 (—) |
| | TMOT w/o Veto | 41.22 (1.74×) | 6.65 | 47.39 (94.78%) | 42.80 (1.66×) | 6.66 | 51.63 (98.55%) | 50.86 (1.89×) | 6.30 | 59.60 (94.90%) |
| | TMOT w/o AESA | 39.81 (1.68×) | 6.69 | 47.78 (95.56%) | 42.02 (1.63×) | 6.69 | 51.54 (98.38%) | 49.51 (1.84×) | 6.19 | 61.60 (98.09%) |
| | TMOT w/o Penalty | 41.88 (1.77×) | 6.68 | 48.78 (97.56%) | 39.62 (1.54×) | 6.59 | 52.24 (99.71%) | 50.99 (1.90×) | 6.26 | 62.00 (98.73%) |
| | Ours | 42.73 (1.81×) | 6.79 | 48.78 (97.56%) | 42.38 (1.65×) | 6.56 | 52.24 (99.71%) | 51.03 (1.90×) | 6.75 | 62.00 (98.73%) |
| Llama-3.2-8B | Baseline | 39.72 (1.00×) | - | 71.34 (—) | 40.74 (1.00×) | - | 77.48 (—) | 43.07 (1.00×) | - | 82.80 (—) |
| | TMOT w/o Veto | 51.58 (1.30×) | 6.71 | 67.68 (94.87%) | 53.13 (1.30×) | 6.23 | 76.04 (98.14%) | 61.51 (1.43×) | 6.62 | 80.80 (97.58%) |
| | TMOT w/o AESA | 53.97 (1.36×) | 6.78 | 70.12 (98.29%) | 52.95 (1.30×) | 6.29 | 74.75 (96.48%) | 58.23 (1.35×) | 6.57 | 79.20 (95.65%) |
| | TMOT w/o Penalty | 49.79 (1.25×) | 6.68 | 69.01 (96.73%) | 52.79 (1.30×) | 6.19 | 76.50 (98.74%) | 59.25 (1.38×) | 6.47 | 81.20 (98.07%) |
| | Ours | 58.23 (1.47×) | 6.57 | 69.51 (97.43%) | 53.09 (1.30×) | 6.19 | 76.50 (98.74%) | 59.79 (1.39×) | 6.75 | 81.20 (98.07%) |
| DSR1-Qwen-7B | Baseline | 23.76 (1.00×) | - | 64.02 (—) | 27.81 (1.00×) | - | 89.99 (—) | 29.23 (1.00×) | - | 86.00 (—) |
| | TMOT w/o Veto | 28.42 (1.20×) | 4.74 | 62.80 (98.09%) | 36.09 (1.30×) | 6.01 | 88.86 (98.74%) | 39.21 (1.34×) | 6.22 | 83.20 (96.74%) |
| | TMOT w/o AESA | 26.53 (1.12×) | 4.70 | 61.59 (96.20%) | 36.12 (1.30×) | 6.10 | 88.17 (97.98%) | 39.09 (1.34×) | 6.20 | 83.60 (97.21%) |
| | TMOT w/o Penalty | 27.41 (1.15×) | 4.64 | 64.02 (100.00%) | 35.98 (1.29×) | 5.85 | 88.65 (98.51%) | 37.46 (1.28×) | 6.12 | 83.60 (97.21%) |
| | Ours | 30.10 (1.27×) | 4.61 | 64.02 (100.00%) | 36.67 (1.32×) | 6.00 | 88.93 (98.82%) | 39.14 (1.34×) | 6.16 | 84.00 (97.67%) |

Removing Veto lowers Pair A's HumanEval recovery from 97.56% to 94.78%. Removing AESA changes its MGSM-EN speedup from $1.90\times$ to $1.84\times$ and recovery from 98.73% to 98.09%. Removing Penalty leaves its HumanEval ACC at 48.78, while throughput changes from 42.73 to 41.88 tokens/s. These results characterize support and calibration components; no w/o-OT or cosine-only condition isolates the necessity or superiority of OT.

### 3.5 RQ4: Hyperparameter Sensitivity

Table 4 gives the full HumanEval parameter sweep for $\tau_{\text{ot}}$, $M$, and $K$, including throughput, MAT, speedup, accuracy, and relative recovery. Defaults remain $K = 3$, $M = 32$, and $\tau_{\text{ot}} = 0.15$.

The support size $K$ directly controls candidate admissibility. At $K = 1$, sharpness equals one and every non-exact proposal fails the support gate before OT, reducing TMOT to exact-match verification. Increasing $K$ enlarges the admissible set but does not yield monotonic quality or throughput gains (Table 4). Entropy normalization retains its $10^{-9}$ stabilizer.

For Llama-3.2-8B, $\tau = 0.2$ yields ACC 70.12 versus 69.51 at the default, with throughput 54.83 versus 58.23 tokens/s. For DSR1-Qwen-7B, $M = 32$ yields 30.10 tokens/s, compared with 27.66 at $M = 8$ and 27.48 at $M = 128$. These results describe the tested settings.

Table 4: HumanEval sensitivity to $\tau = \tau_{\text{ot}}$, $M$, and $K$. Rate is tokens/s (speedup); ACC includes relative recovery. Defaults are marked Ours. At $K = 1$, the support gate rejects every non-exact proposal before OT.

| | | Llama-3.1-13B | | | Llama-3.2-8B | | | DSR1-Qwen-7B | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Param** | **Value** | **Rate (×)** | **MAT** | **ACC (Rec)** | **Rate (×)** | **MAT** | **ACC (Rec)** | **Rate (×)** | **MAT** | **ACC (Rec)** |
| $\tau$ | 0.05 | 40.84 (1.73×) | 6.55 | 48.17 (96.34%) | 49.59 (1.25×) | 6.43 | 68.29 (95.72%) | 27.72 (1.17×) | 4.47 | 63.41 (99.05%) |
| | 0.1 | 41.00 (1.73×) | 6.59 | 48.17 (96.34%) | 57.07 (1.44×) | 6.58 | 68.90 (96.58%) | 28.35 (1.19×) | 4.53 | 62.80 (98.09%) |
| | 0.15 (Ours) | 42.73 (1.81×) | 6.79 | 48.78 (97.56%) | 58.23 (1.47×) | 6.57 | 69.51 (97.43%) | 30.10 (1.27×) | 4.61 | 64.02 (100.00%) |
| | 0.2 | 41.84 (1.77×) | 6.69 | 48.78 (97.56%) | 54.83 (1.38×) | 6.81 | 70.12 (98.29%) | 28.87 (1.22×) | 4.65 | 63.41 (99.05%) |
| | 0.25 | 38.89 (1.64×) | 6.72 | 48.17 (96.34%) | 54.67 (1.38×) | 6.83 | 69.51 (97.43%) | 27.05 (1.14×) | 4.69 | 63.41 (99.05%) |
| $M$ | 8 | 37.86 (1.60×) | 6.62 | 48.78 (97.56%) | 57.82 (1.46×) | 6.54 | 69.51 (97.43%) | 27.66 (1.16×) | 4.64 | 62.80 (98.09%) |
| | 16 | 41.84 (1.77×) | 6.59 | 48.78 (97.56%) | 58.19 (1.47×) | 6.63 | 70.12 (98.29%) | 29.27 (1.23×) | 4.66 | 63.41 (99.05%) |
| | 32 (Ours) | 42.73 (1.81×) | 6.79 | 48.78 (97.56%) | 58.23 (1.47×) | 6.57 | 69.51 (97.43%) | 30.10 (1.27×) | 4.61 | 64.02 (100.00%) |
| | 64 | 41.07 (1.74×) | 6.67 | 48.17 (96.34%) | 54.72 (1.38×) | 6.71 | 69.51 (97.43%) | 28.50 (1.20×) | 4.64 | 64.02 (100.00%) |
| | 128 | 41.20 (1.74×) | 6.66 | 48.78 (97.56%) | 52.46 (1.32×) | 6.72 | 68.90 (96.58%) | 27.48 (1.16×) | 4.64 | 63.41 (99.05%) |
| $K$ | 1 | 38.27 (1.62×) | 6.35 | 46.95 (93.90%) | 53.16 (1.34×) | 6.20 | 70.73 (99.14%) | 27.01 (1.14×) | 4.35 | 64.02 (100.00%) |
| | 3 (Ours) | 42.73 (1.81×) | 6.79 | 48.78 (97.56%) | 58.23 (1.47×) | 6.57 | 69.51 (97.43%) | 30.10 (1.27×) | 4.61 | 64.02 (100.00%) |
| | 5 | 42.45 (1.79×) | 6.71 | 48.17 (96.34%) | 57.86 (1.46×) | 6.68 | 67.68 (94.87%) | 28.44 (1.20×) | 4.71 | 63.41 (99.05%) |
| | 10 | 41.07 (1.74×) | 6.67 | 48.78 (97.56%) | 57.23 (1.44×) | 6.61 | 68.29 (95.72%) | 26.17 (1.10×) | 4.63 | 62.20 (97.16%) |

## 4 Related Work

**Distribution-preserving speculative decoding.** Speculative sampling preserves the target distribution through verification and correction (Leviathan et al., 2023), while greedy blockwise decoding has a distinct acceptance rule (Stern et al., 2018). Medusa and EAGLE improve drafting (Cai et al., 2024; Li et al., 2024); PEARL overlaps execution and adapts draft length (Liu et al., 2025). SpecTr already uses OT for draft selection with target-distribution fidelity (Sun et al., 2023). OT itself is therefore not new in speculative decoding.

**Relaxed and confidence-aware verification.** FSD uses distributional divergences (Holsman et al., 2025); FLy uses target entropy and deferred verification (Li et al., 2026). Judge Decoding trains a compact judge on target embeddings (Bachmann et al., 2025), while MARS relaxes verification according to target-logit decision stability (Song et al., 2026). Analysis of lossy verification connects relaxation to altered output distributions and quality failure modes (Wang et al., 2026). Entropy-aware relaxation is not unique to TMOT; its uncertainty signals calibrate a geometric state gate alongside candidate probability constraints.

**Semantic and representation-based verification.** SemanticSpec probes internal states to verify semantic sequences (Dong et al., 2026); Reflective Verification combines original and prompt-probed reflective distributions (Wang et al., 2025). Semantic relaxation is thus not unique to TMOT. Representation learning and geometry provide broader context (BehnamGhader et al., 2024; Park et al., 2023), but static token proximity does not ensure contextual correctness. TMOT specifically uses target representations as the ground geometry for predictive probability measures, computed with entropy-regularized transport (Cuturi, 2013; Peyré & Cuturi, 2019).

## 5 Conclusion

TMOT formulates speculative verification as probability-measure alignment: binary token identity yields total variation, and target representations induce graded ground geometry. Its factorized verifier combines candidate admissibility with uncertainty-adaptive predictive-state compatibility. Experiments yield up to $1.90\times$ speedup and a small verification-time share, with task-dependent accuracy recovery.

**Limitations.** Static token geometry does not capture all contextual semantics. TMOT is a lossy greedy verifier, with no exact distribution-preservation or semantic-correctness guarantee; open-ended, high-temperature generation is untested. Its OT score is distribution-level rather than candidate-conditioned, and the ablations do not isolate OT from a w/o-OT or cosine-only alternative.

### Reproducibility Statement

The algorithm, hyperparameters, evaluation metrics, hardware, and decoding configurations are described in Sections 2–3. Exact model checkpoints and implementation details are provided in Appendix B, complete proofs of the theoretical claims are included in Appendix A, and hyperparameter sensitivity is reported in Table 4.

### AI Use Statement

Generative AI tools were used to assist with English translation and language editing, structural revision, conceptual framing and methodological feedback, consistency checking, formulation and development of mathematical claims and proofs, interpretation and presentation of existing experimental results, literature discovery and reference checking, figure preparation, and LaTeX editing. No experimental data were generated by AI tools. All AI-assisted content was reviewed by the authors, who take responsibility for the final manuscript, including its technical claims, derivations, experimental results, model configurations, and AI-assisted content.

## References

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

Gregor Bachmann, Sotiris Anagnostidis, Albert Pumarola, Markos Georgopoulos, Artsiom Sanakoyeu, Yuming Du, Edgar Schönfeld, Ali Thabet, and Jonas Kohler. Judge decoding: Faster speculative sampling requires going beyond model alignment. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mtSSFiqW6y.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024.

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

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Ximing Dong, Shaowei Wang, Dayi Lin, Boyuan Chen, and Ahmed E. Hassan. Beyond tokens: Semantic-aware speculative decoding for efficient inference by probing internal states. *arXiv preprint arXiv:2602.03708*, 2026. URL https://arxiv.org/abs/2602.03708.

Maximilian Holsman, Yukun Huang, and Bhuwan Dhingra. Fuzzy speculative decoding for a tunable accuracy-runtime tradeoff. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26257–26273, 2025.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International conference on machine learning*, pp. 19274–19286. PMLR, 2023.

Jinze Li, Yixing Xu, Guanchen Li, Shuo Yang, Jinfeng Xu, Xuanwu Yin, Dong Li, Edith Ngai, and Emad Barsoum. Training-free loosely speculative decoding: Accepting semantically correct drafts beyond exact match. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.22972.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*, 2024.

Tianyu Liu, Yun Li, Qitan Lv, Kai Liu, Jianchen Zhu, Winston Hu, and Xiao Sun. Pearl: Parallel speculative decoding with adaptive draft length. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.11850.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073. URL https://www.nowpublishers.com/article/Details/MAL-073.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022. URL https://arxiv.org/abs/2210.03057.

Jingwei Song, Xinyu Wang, Hanbin Wang, Xiaoxuan Lei, Tianyu Shi, Shixin Han, Eric Yang, Xiao-Wen Chang, and Lynn Ai. MARS: Unleashing the power of speculative decoding via margin-aware verification. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8652–8663. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.421. URL https://aclanthology.org/2026.findings-acl.421/.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31, 2018.

Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix Yu. SpecTr: Fast speculative decoding via optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2310.15141.

Tianyu Wang, Yuxuan Zhou, Heng Li, Wenbin Wang, Zikai Xiao, Chunrui Zheng, and Junyuan Shang. Revisiting lossy verification in speculative decoding: Mechanisms, trade-offs, and failure modes. *arXiv preprint arXiv:2607.26627*, 2026. URL https://arxiv.org/abs/2607.26627.

Yixuan Wang, Yijun Liu, Shiyu Ji, Yuzhuang Xu, Yang Xu, Qingfu Zhu, and Wanxiang Che. Think before you accept: Semantic reflective verification for faster speculative decoding. *arXiv preprint arXiv:2505.18629*, 2025. URL https://arxiv.org/abs/2505.18629.

# A PROOFS

## A.1 PROOF OF PROPOSITION 1

*Proof.* For any coupling $T$, $\langle T, C_{\text{disc}} \rangle_F = 1 - \sum_x T_{xx}$. Its marginals imply $T_{xx} \leq \min\{\mu_d(x), \mu_t(x)\}$, giving the stated lower bound. To attain it, put $m_x = \min\{\mu_d(x), \mu_t(x)\}$ on the diagonal. The residual masses $r_x = \mu_d(x) - m_x$ and $s_x = \mu_t(x) - m_x$ are nonnegative, have disjoint supports, and sum to the same $q = 1 - \sum_x m_x$. If $q = 0$, the diagonal plan is sufficient. Otherwise add the residual coupling $r_x s_y / q$. It has the required residual marginals and zero diagonal, so its cost is $q$. Finally, $|\mu_d(x) - \mu_t(x)| = \mu_d(x) + \mu_t(x) - 2m_x$; summing and dividing by two gives $q = \text{TV}(\mu_d, \mu_t)$. □

## A.2 PROOF OF PROPOSITION 2

*Proof.* Cauchy–Schwarz gives $-1 \leq \langle \bar{\mathbf{e}}(x), \bar{\mathbf{e}}(y) \rangle \leq 1$, so the cost lies in $[0, 1]$. For $x = y$, the inner product is one and the cost is zero. Thus the cost is bounded above by $C_{\text{disc}}$ at every pair. The finite-support coupling polytope is nonempty and compact, so a minimizer $T^*_{\text{disc}}$ exists. Using the same coupling for the geometric objective yields

$$W_{C_{\text{TMOT}}} \leq \langle T^*_{\text{disc}}, C_{\text{TMOT}} \rangle_F \leq \langle T^*_{\text{disc}}, C_{\text{disc}} \rangle_F = W_{C_{\text{disc}}} = \text{TV}.$$

This argument assumes exact marginal feasibility and optimization, not the finite-step score in equation 10. □

## A.3 PROOF OF PROPOSITION 3

*Proof.* In the non-exact branch of equation 5, acceptance requires $G_i R_i = 1$. Both factors are indicators, so $G_i = 1$. Every indicator in equation 12 must then equal one. The first two give Top-$K$ membership and the absolute probability threshold. If $\mathcal{S}_i > \tau_{\text{sharp}}$, the remaining disjunction can hold only through $P_T^{(i)}(x_i^d) \geq \rho p_{i,\max}^t$. No assertion is made about these conditions for the direct exact-match branch. □

## A.4 PROOF OF PROPOSITION 4

*Proof.* Because $\Phi_i > 0$, dividing equation 14 by $\Phi_i$ gives equation 15 with the same strict inequality. For $H < \theta$, $\partial B / \partial H = \gamma B \geq 0$; for $H > \theta$, it is zero. The function is continuous at $\theta$, with nonnegative one-sided derivatives. For all $H$,

$$\frac{\partial B}{\partial S} = -\kappa \tau_{\text{ot}} S \exp[-\gamma \max(0, \theta - H)] \leq 0.$$

These establish the coordinatewise monotonicities on the stated domain. □

# B IMPLEMENTATION DETAILS

**Model checkpoints.** Pairs A and B use `meta-llama/Llama-3.2-3B-Instruct` as the drafter, with targets `TroyDoesAI/Llama-3.1-13B-Instruct` and `theo77186/Llama-3.2-8B-Instruct`, respectively. Pair C uses `deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B` and `deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`.

**Mean Advanced Tokens.** The implementation computes MAT as `self.custom_total_toks` divided by `self.custom_draft_rounds`. Each round adds the accepted draft tokens and one target-generated token to the numerator: a bonus token when the full block is accepted, or a correction/recovery token at the first rejection. With batch size one, the latter contribution is implemented as `+ batch_size`. Hence $\text{MAT} = R^{-1} \sum_{r=1}^{R} (A_r + 1)$. TMOT uses $N = 8$, so $0 \leq A_r \leq 8$ and $1 \leq \text{MAT} \leq 9$. This proposal length is distinct from the $L = 5$ Sinkhorn iterations.

**Scaled raw entropy.** For retained mass $m = \sum_{x \in \mathrm{TopJ}} P_T(x) \leq 1$ and $q_x = P_T(x)/m$, the raw entropy is $-m \log m + mH(q) \leq m \log(J/m)$. For $J \geq 3$, the right-hand side increases on $0 < m \leq 1$, since its derivative is $\log(J/m) - 1 > 0$, and is at most $\log J$. This justifies the range used for default $K = 3, M = 32$ without changing the computation. The implementation divides by $\log J + 10^{-9}$. At $K = 1$, the denominator remains finite, while the candidate gate prevents every non-exact proposal from reaching OT (Section 3.5).

**Numerical implementation.** Target embeddings are normalized using `F.normalize(emb_manifold, p=2, dim=-1)`, with default `eps=1e-12`. Propositions 1–2 concern ideal probability measures and, for the geometric cost, nonzero unit-normalized embeddings. Union indices select original full-distribution probabilities before independent renormalization; they do not zero-fill the separate Top-$M$ lists. The normalization in equation 6 defines ideal probability measures; the implementation additionally stabilizes the draft normalization denominator by $10^{-8}$. The practical score uses the other stabilizers and fixed iteration count specified in Section 2.5.

**Vectorized verification and recovery.** Position-wise Boolean gates are evaluated in parallel. The cumulative minimum of the acceptance mask retains only the accepted prefix. In zero-based code, the first rejected position is recovered inside `_verify_tokens` by `recovered_tok = target_tokens[b, num_accepted]`, where `target_tokens = target_probs_k.argmax(dim=-1)`. This is a target greedy token, not a sample from a residual or target distribution.