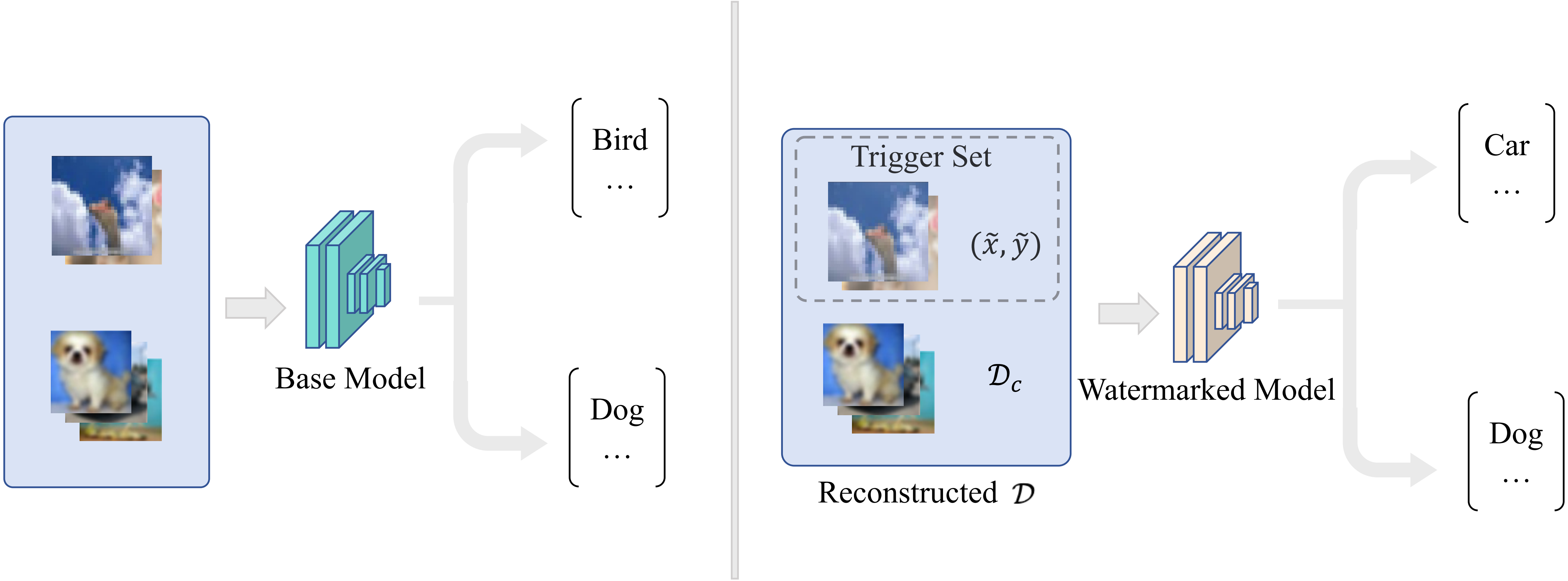
Dataset Reduction and Watermark Removal via Self-supervised Learning for Model Extraction Attack

Hao Luan[1], Xue Tan[1], Zhiheng Li[3], Jun Dai[2], Xiaoyan Sun[2] and Ping Chen[1]

[1]Fudan University, [2]Worcester Polytechnic Institute, [3]Shandong University,

Background: Black-box Model Watermarking



Base Model

Bird …

Dog …

Trigger Set

$(\tilde{x}, \tilde{y})$

$\mathcal{D}_c$

Reconstructed $\mathcal{D}$

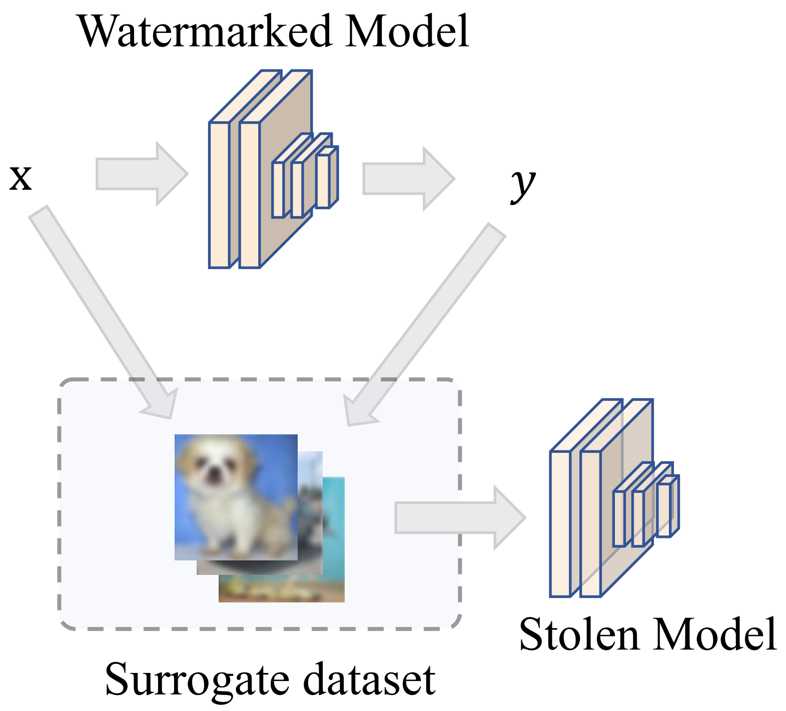
Watermarked Model

Car …

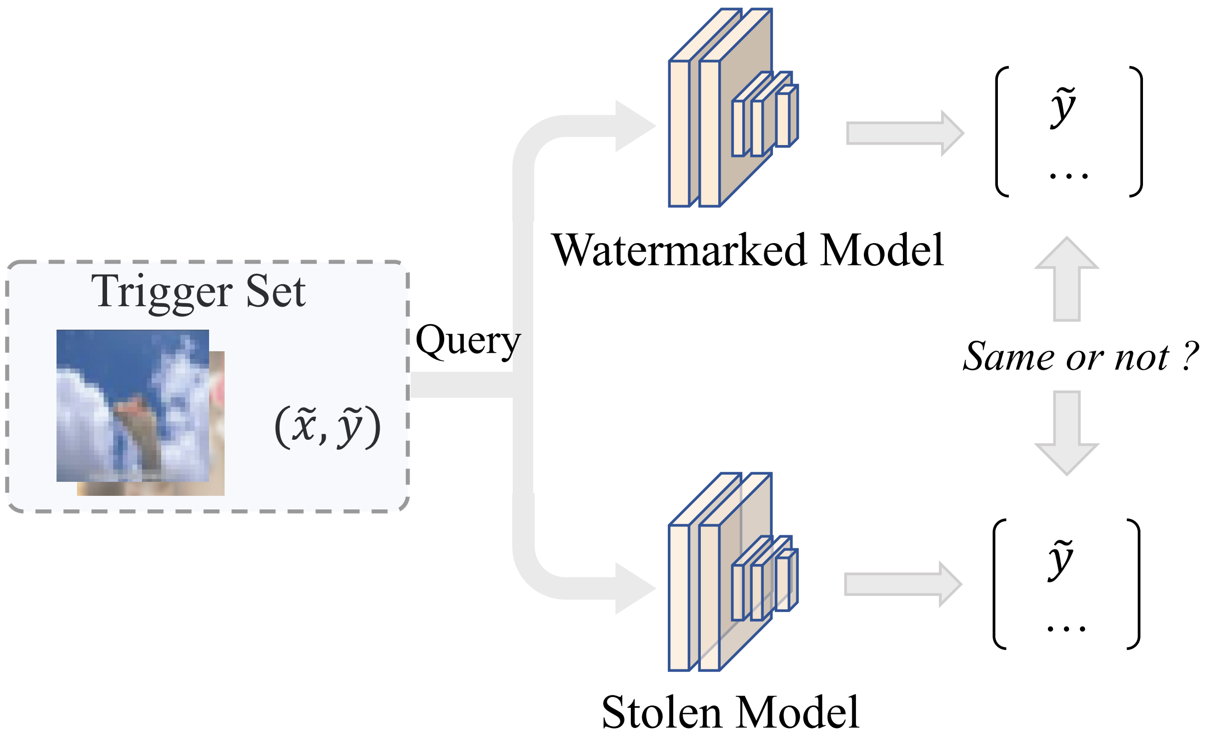
Dog …

Background: Model Extraction Attack

Training Stolen Model:



Verify:



Motivation & Challenges

Challenge 1:
Inefficiency in Query Usage

- Standard attacks rely on inefficient random or boundary sampling.
- Result: Require massive queries (e.g., >10k) to achieve competitive accuracy.

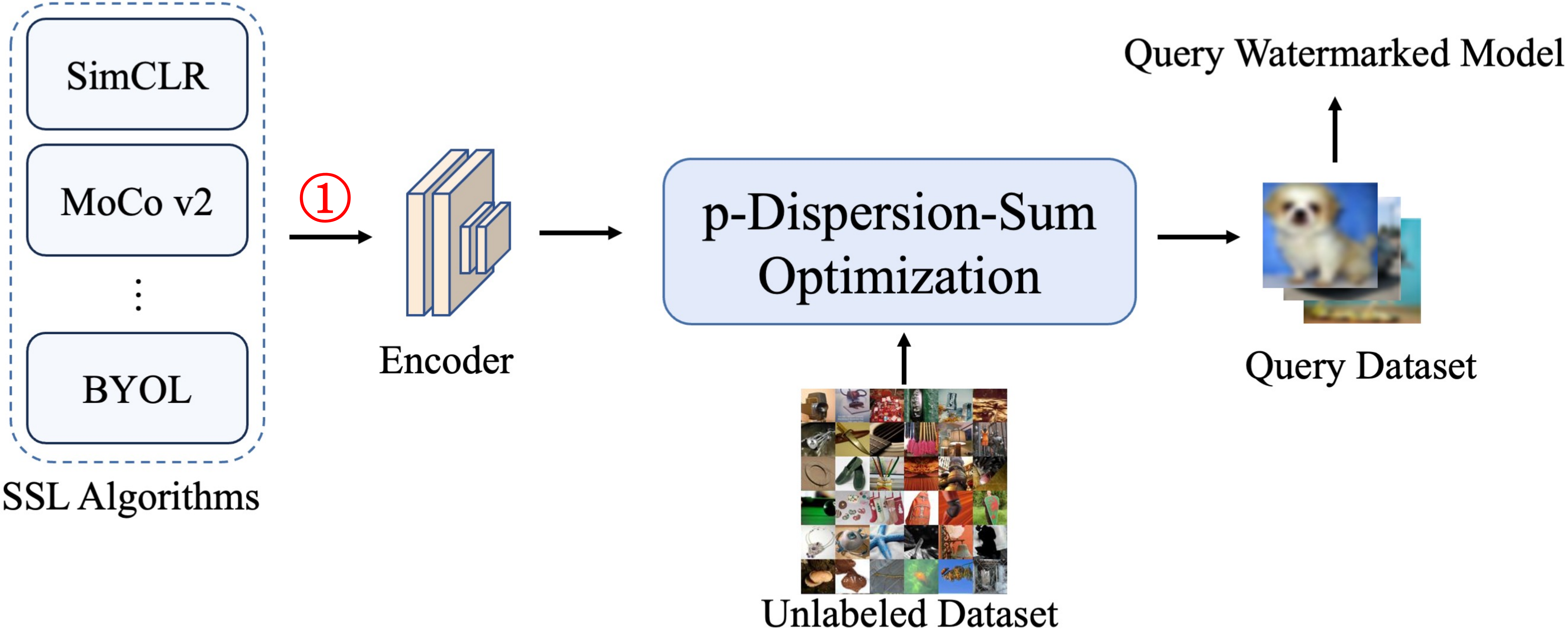
Challenge 2:
Ineffective Watermark Removal

- Existing data reduction methods operate in the pixel space.
- Result: Fail to distinguish triggers, inevitably learning the watermark.

Our Goal:

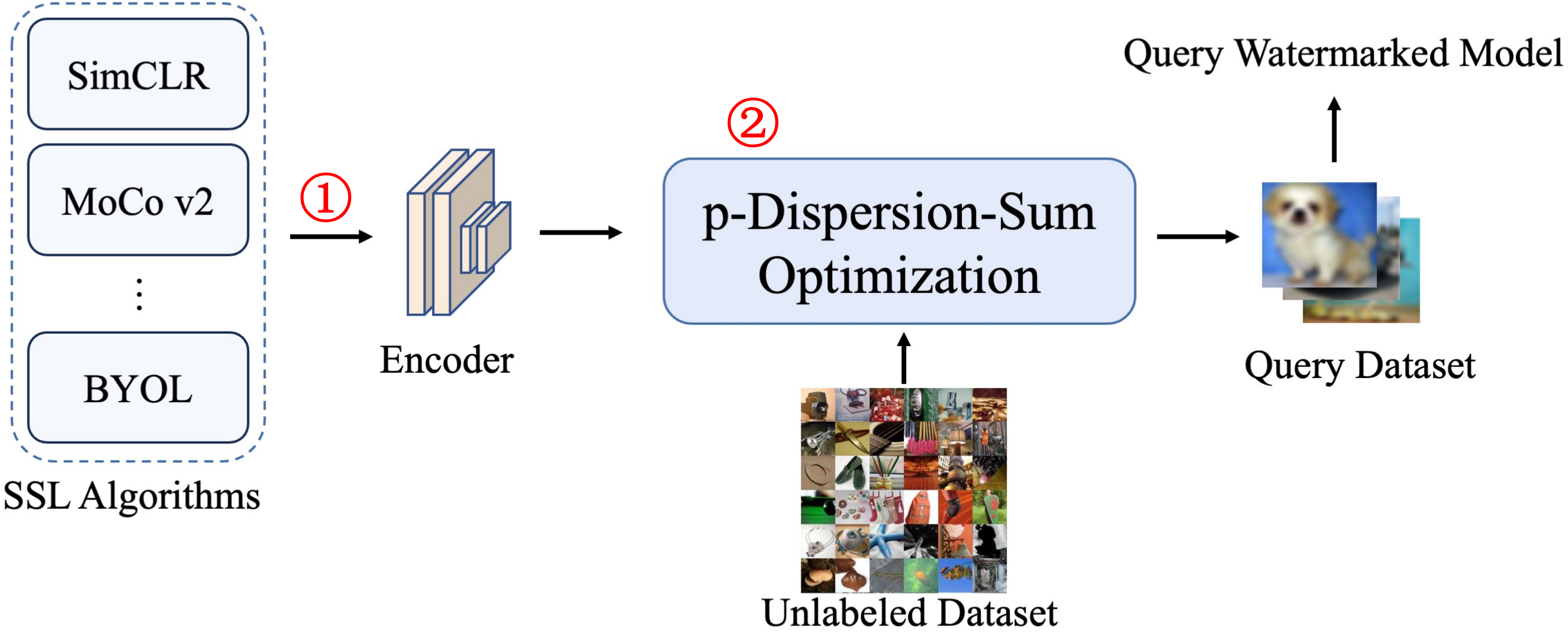
High Accuracy Extraction with
Minimal Queries & Removed Watermarks

Method Overview



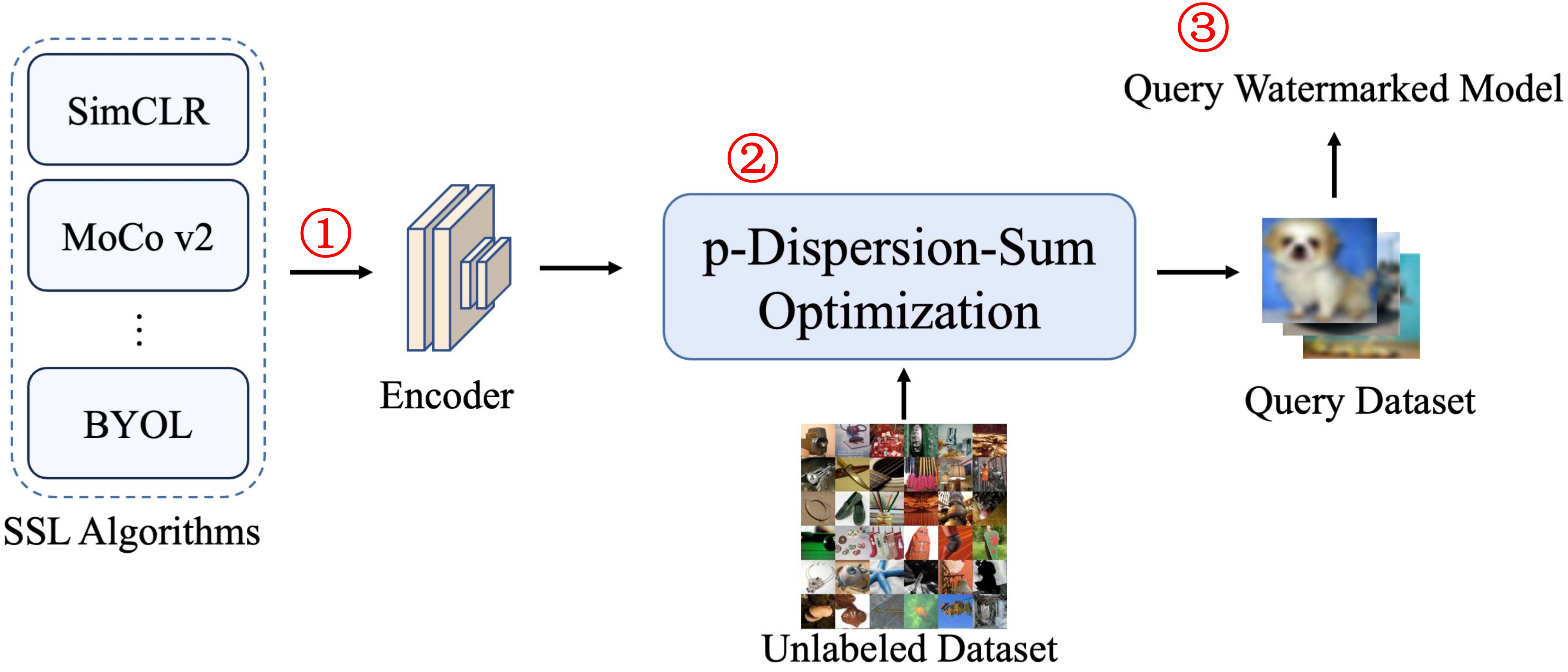
- **Step ①: Feature Extraction:** Train an encoder via SSL to capture intrinsic, watermark-agnostic representations.

Method Overview



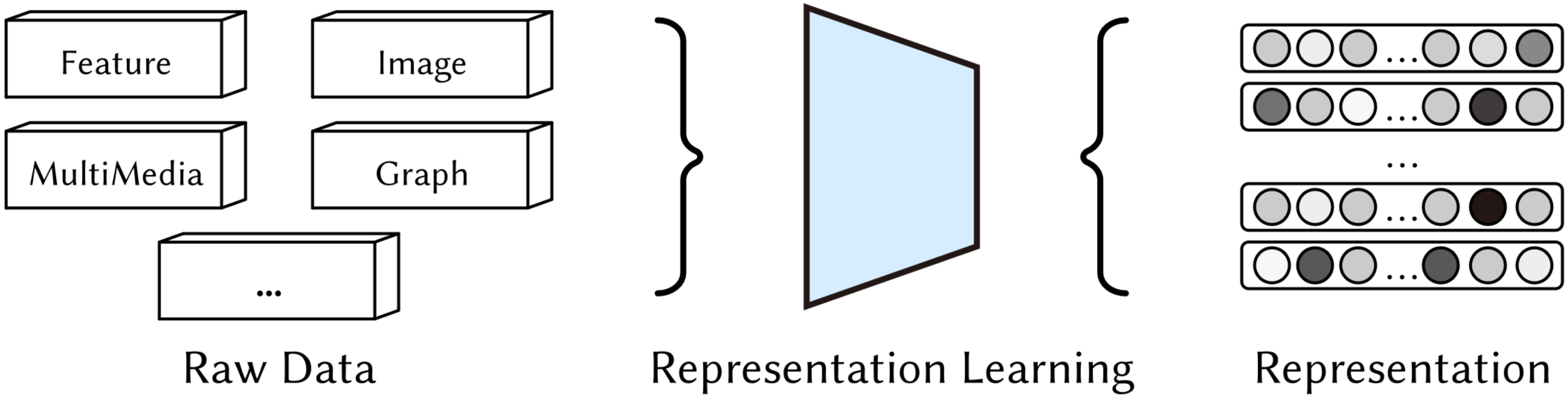
- **Step ①: Feature Extraction:** Train an encoder via SSL to capture intrinsic, watermark-agnostic representations.
- **Step ②: Query Selection:** Select diverse queries by solving the p-dispersion-sum problem.

Method Overview



- **Step ①: Feature Extraction:** Train an encoder via SSL to capture intrinsic, watermark-agnostic representations.
- **Step ②: Query Selection:** Select diverse queries by solving the p-dispersion-sum problem.
- **Step ③: Model Extraction:** Query the victim model with selected samples to train a clean surrogate.

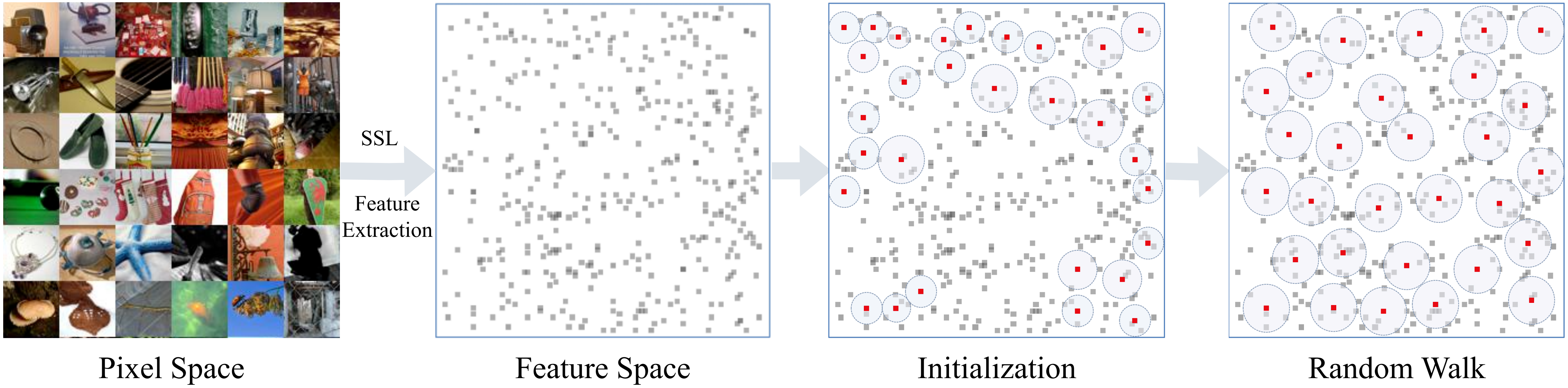
Step 1: Feature Extraction via SSL



- **Feature Extraction:** Transforms raw pixels into **intrinsic, semantic representations** using unlabeled data.

- **Watermark Removal:** SSL ignores artificial, label-bound patterns. Triggers are naturally isolated as **outliers** in the feature space.

Step 2: Data Reduction



- **Problem Formulation:** We formulate query selection as a **p-dispersion-sum problem** to maximize the distance and diversity among selected features.
- **Greedy Initialization:** We construct an initial query subset by greedily selecting features with the maximum distance to already chosen ones.
- **Iterative Random Walk:** We iteratively add random features and drop the least contributing ones, ultimately achieving a **uniform, broad coverage** of the entire feature space.

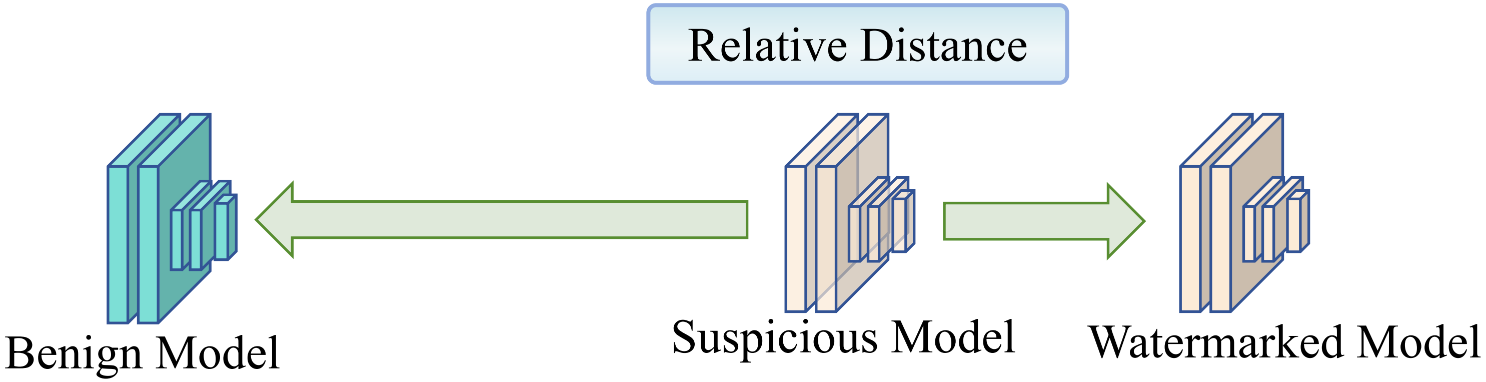
Ownership Verification: The Relative Distance Ratio r

Limitations of Prior Metrics

- **WSR**
  - Focus only on the **absolute distance** to the watermarked model.
  - *Issue:* Prone to false claims if benign models unintentionally match triggers.
- **Threshold-based Approaches**
  - Consider benign models, but rely on **manually predefined, linear** transformations.
  - *Issue:* Heuristic and lacks statistical robustness.

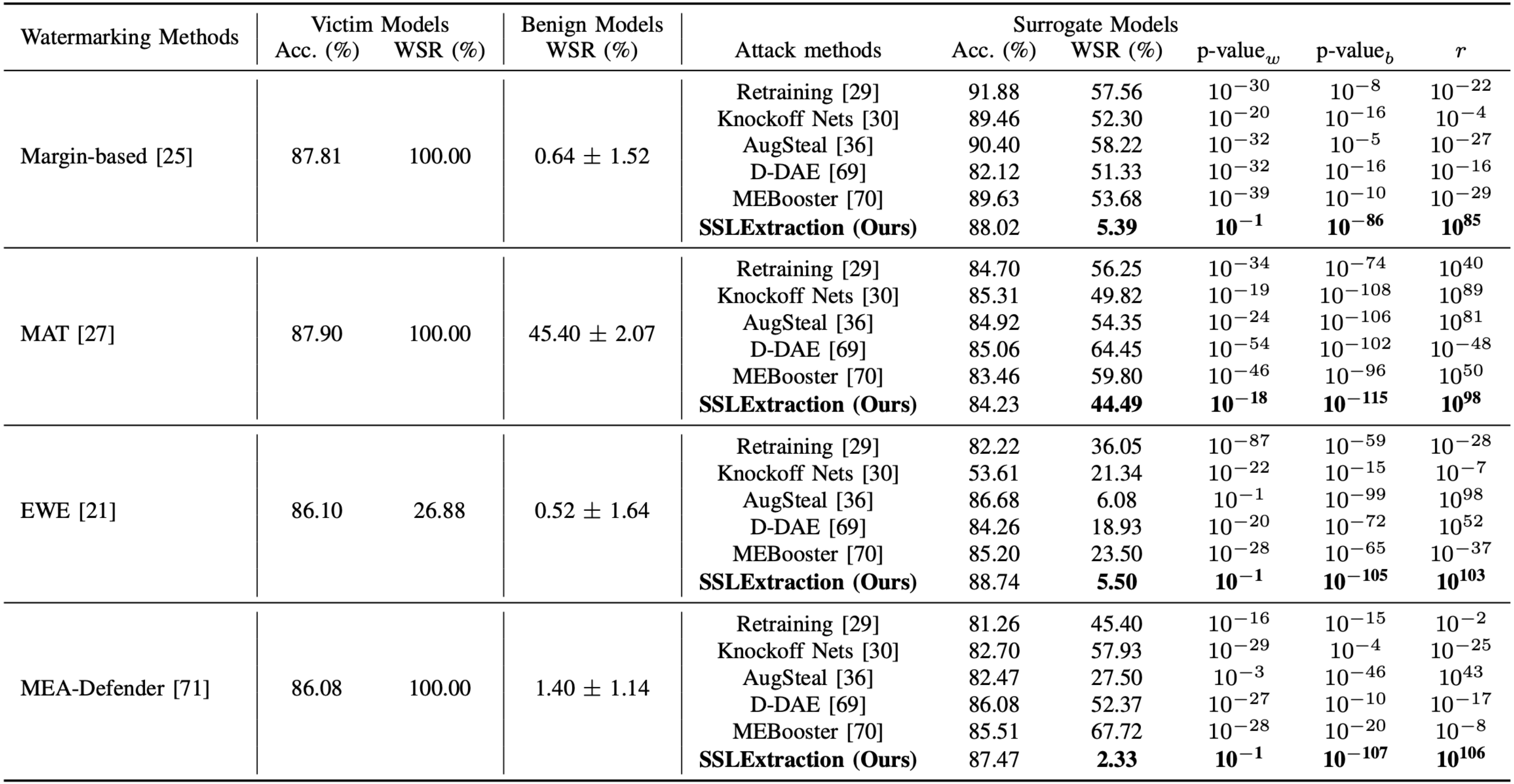
Our Proposed Metric: Ratio r

$$r = \frac{\text{p-value}_w}{\text{p-value}_b}$$



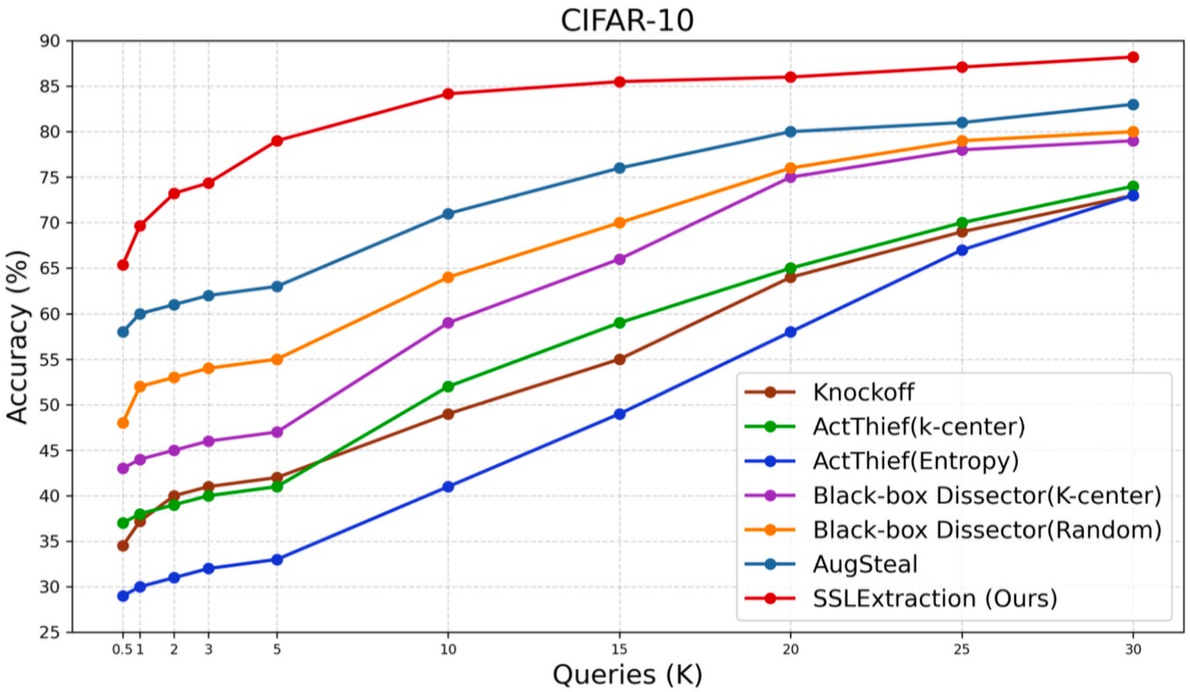
Main Results: Watermark Removal

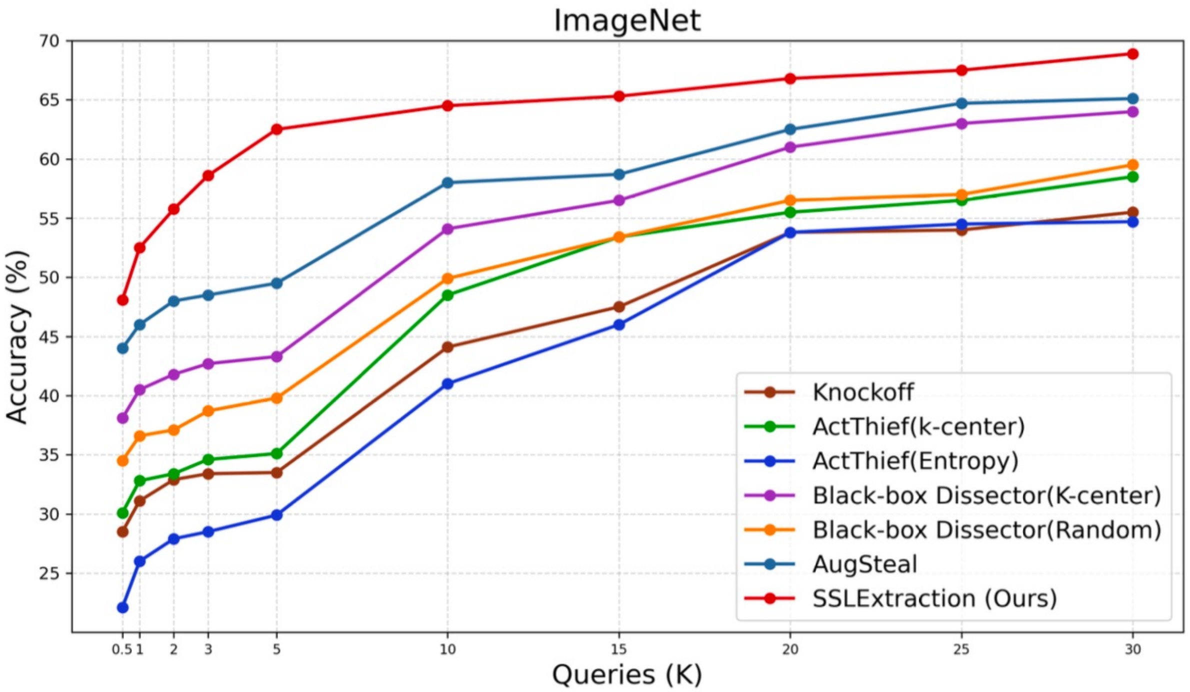
TABLE I: Results for model extraction attacks against watermarking schemes on CIFAR-10 dataset.



Watermarking Methods	Victim Models		Benign Models	Surrogate Models					
	Acc. (%)	WSR (%)	WSR (%)	Attack methods	Acc. (%)	WSR (%)	p-value$_w$	p-value$_b$	r
Margin-based [25]	87.81	100.00	0.64 ± 1.52	Retraining [29]	91.88	57.56	10^{-30}	10^{-8}	10^{-22}
				Knockoff Nets [30]	89.46	52.30	10^{-20}	10^{-16}	10^{-4}
				AugSteal [36]	90.40	58.22	10^{-32}	10^{-5}	10^{-27}
				D-DAE [69]	82.12	51.33	10^{-32}	10^{-16}	10^{-16}
				MEBooster [70]	89.63	53.68	10^{-39}	10^{-10}	10^{-29}
				SSLExtraction (Ours)	88.02	**5.39**	$\mathbf{10^{-1}}$	$\mathbf{10^{-86}}$	$\mathbf{10^{85}}$
MAT [27]	87.90	100.00	45.40 ± 2.07	Retraining [29]	84.70	56.25	10^{-34}	10^{-74}	10^{40}
				Knockoff Nets [30]	85.31	49.82	10^{-19}	10^{-108}	10^{89}
				AugSteal [36]	84.92	54.35	10^{-24}	10^{-106}	10^{81}
				D-DAE [69]	85.06	64.45	10^{-54}	10^{-102}	10^{-48}
				MEBooster [70]	83.46	59.80	10^{-46}	10^{-96}	10^{50}
				SSLExtraction (Ours)	84.23	**44.49**	$\mathbf{10^{-18}}$	$\mathbf{10^{-115}}$	$\mathbf{10^{98}}$
EWE [21]	86.10	26.88	0.52 ± 1.64	Retraining [29]	82.22	36.05	10^{-87}	10^{-59}	10^{-28}
				Knockoff Nets [30]	53.61	21.34	10^{-22}	10^{-15}	10^{-7}
				AugSteal [36]	86.68	6.08	10^{-1}	10^{-99}	10^{98}
				D-DAE [69]	84.26	18.93	10^{-20}	10^{-72}	10^{52}
				MEBooster [70]	85.20	23.50	10^{-28}	10^{-65}	10^{-37}
				SSLExtraction (Ours)	88.74	**5.50**	$\mathbf{10^{-1}}$	$\mathbf{10^{-105}}$	$\mathbf{10^{103}}$
MEA-Defender [71]	86.08	100.00	1.40 ± 1.14	Retraining [29]	81.26	45.40	10^{-16}	10^{-15}	10^{-2}
				Knockoff Nets [30]	82.70	57.93	10^{-29}	10^{-4}	10^{-25}
				AugSteal [36]	82.47	27.50	10^{-3}	10^{-46}	10^{43}
				D-DAE [69]	86.08	52.37	10^{-27}	10^{-10}	10^{-17}
				MEBooster [70]	85.51	67.72	10^{-28}	10^{-20}	10^{-8}
				SSLExtraction (Ours)	87.47	**2.33**	$\mathbf{10^{-1}}$	$\mathbf{10^{-107}}$	$\mathbf{10^{106}}$

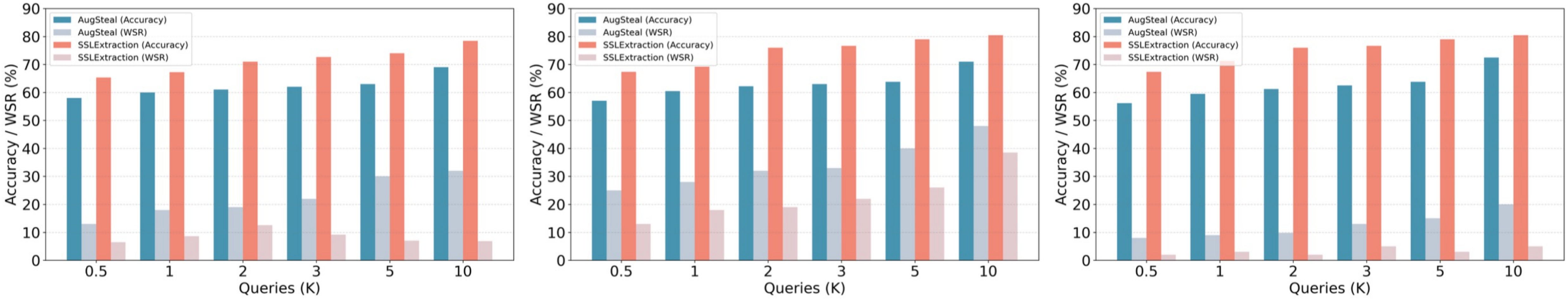
Main Results: Query Efficiency





Our method consistently achieves **high accuracy** under **extremely limited query budgets.**

Main Results: Simultaneous Data Reduction & Watermark Removal



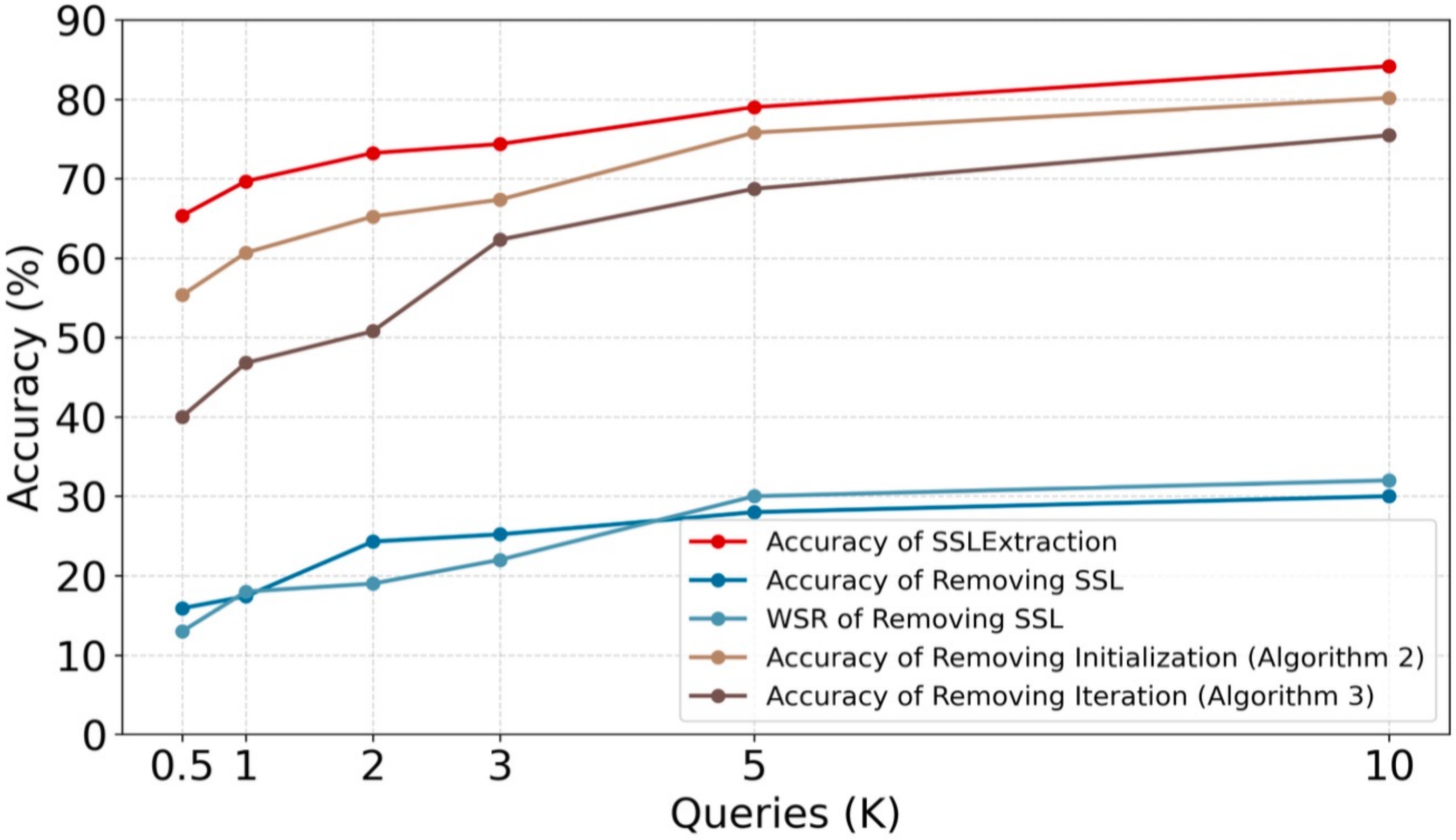
(a) Margin-based Watermarking

(b) MAT

(c) MEA-Defender

Our method maintains **competitive accuracy** while strictly suppressing the **WSR to a near-zero level** across all budgets.

Ablation Study



- **Without SSL:** Accuracy collapses and watermark is retained. (SSL is fundamental for isolating triggers).
- **Without Greedy Initialization:** Noticeable performance drop. (Highlights the need for a strong start).
- **Without Random Walk Iteration:** Accuracy degrades significantly. (Essential for uniform space coverage).

THANKS