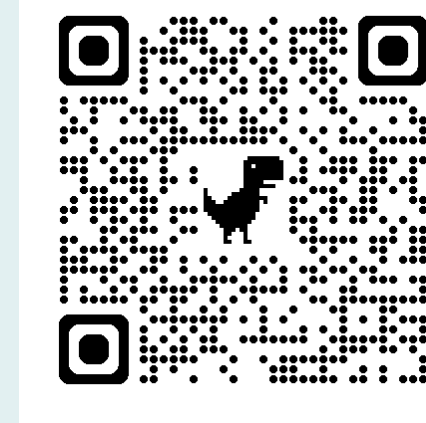


採択されるデータセット論文を書くために

村上尚生(中京大), 馬上凌(東大), 原涼成(慶大), 大石康太郎(慶大), 丸山虹樹(個人)



概要

調査目的：トップカンファレンスで発表されたデータセット論文を分析し、採択される論文の特徴を明らかにする。

調査方法：OpenReview で評価された項目を集計するとともに、論文を「構築前」「構築中」「構築後」の3つの観点から分析し、各段階で重視される記載内容を整理した。

調査結果：データセットの新規性だけでなく、**アノテーション品質**や**評価実験の厳密さ**が採否を決める要因である。

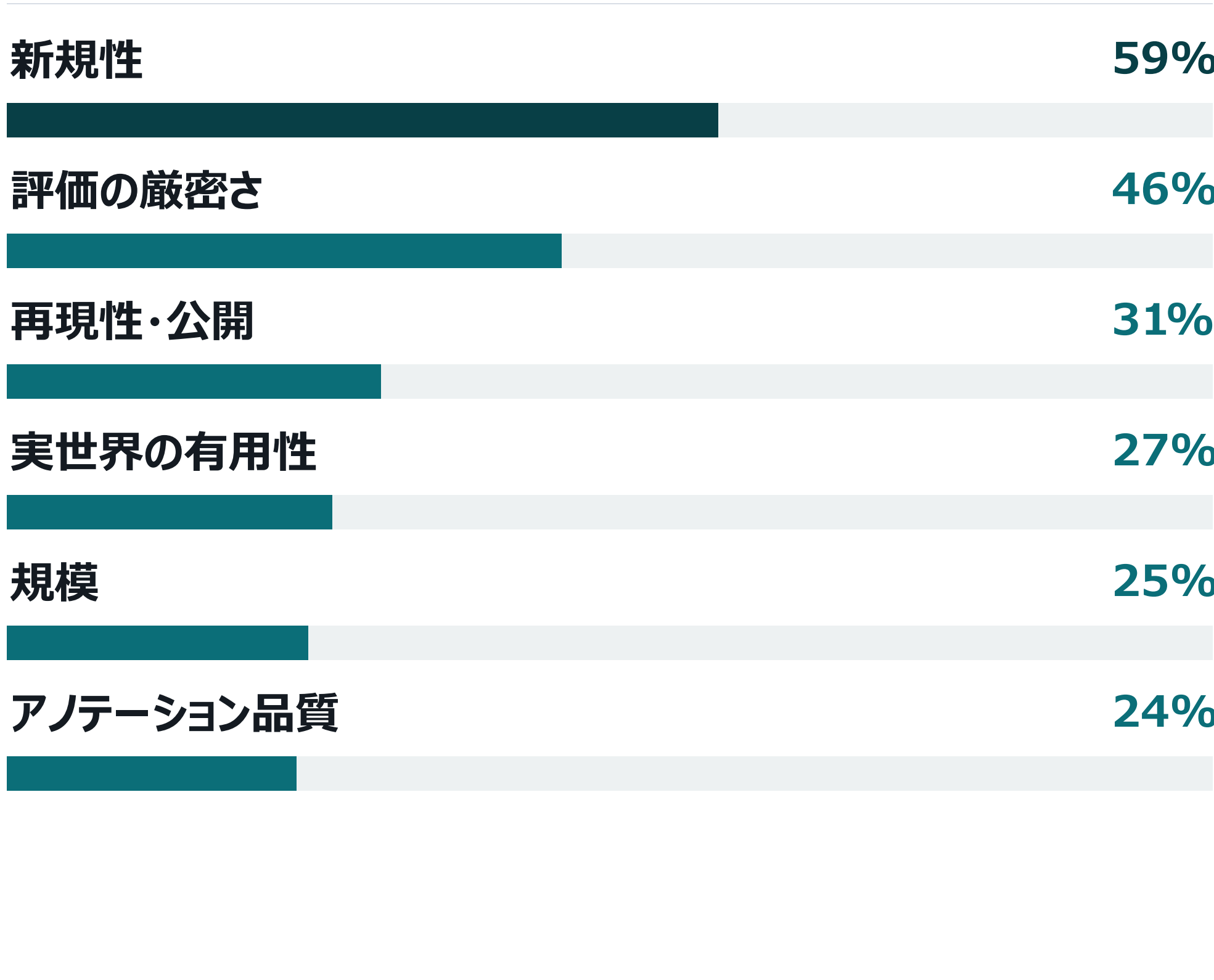
データセット論文は「何で」通るのか

NeurIPS DATASETS & BENCHMARKS TRACK · 2024–2025 OpenReview分析

1,117 投稿 | 956 採択全数 | 4,246 査読

1 採択の決め手

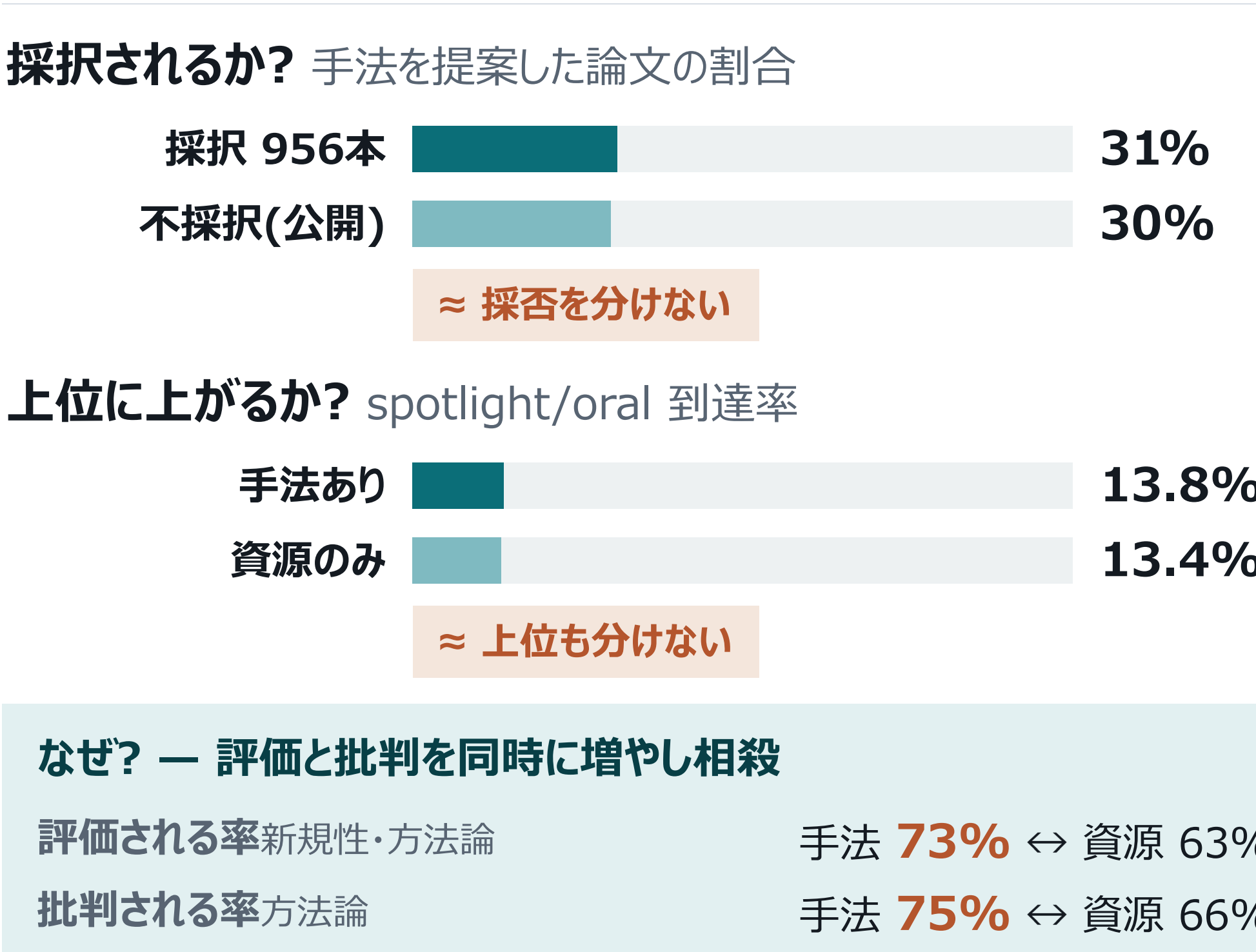
新規性 “だけ” では通らない



採択956本で査読者に称賛された割合。
新規性を土台に、**厳密な評価・公開・規模**を上乗せする。

2 データのみ / 手法も

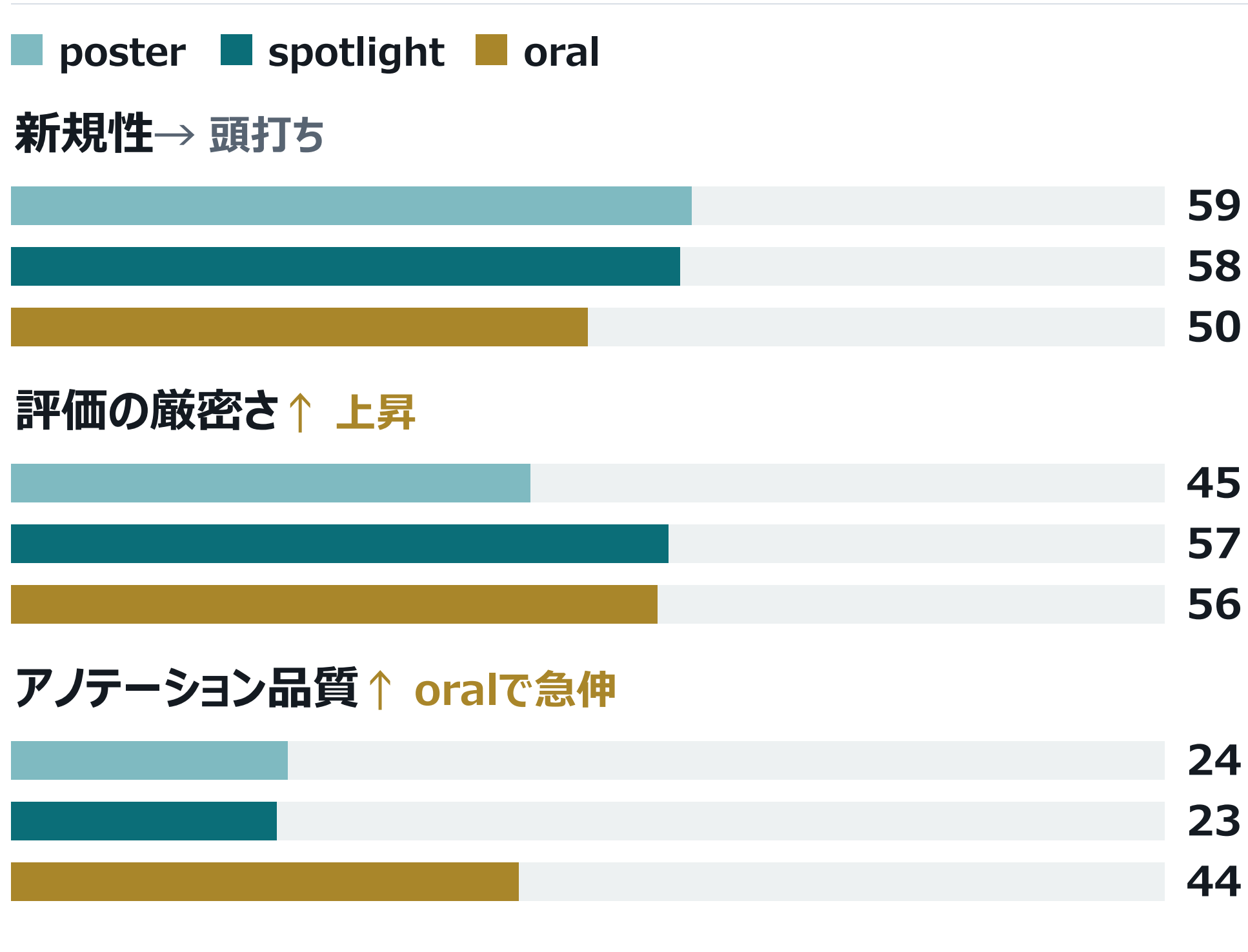
手法を足しても “得” にならない



採択の**31%**が新手法も提案(うち手法が主貢献なのは**10%**、多くは補助的)。それでも採否・tierは動かない。

3 POSTER → SPOTLIGHT / ORAL

上位は “作り込みの深さ” で差がつく

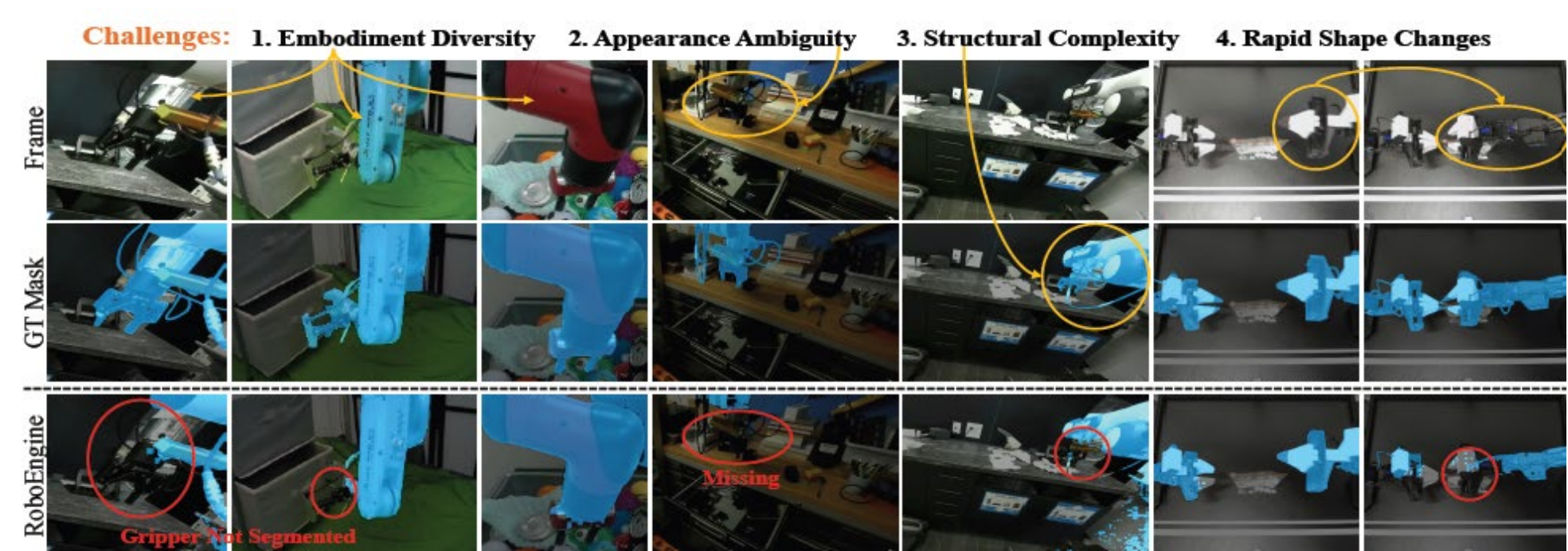


各バーは、その tier 内で査読者に称賛された割合。**新規性は全 tier で高い**が**上位で伸びず頭打ち**(59→50%)。上位で伸びるのは**評価の厳密さ**(45→57%)と、oralで急伸する**アノテーション品質**(24→44%)。

データセット構築前：データセットの必要性を明確化

従来手法の性能的な観点

ロボットセグメンテーションのための動画像データセットの必要性を主張



RobotSeg [Haiyang+, CVPR2026]

データセット構築の観点

従来データセットにおけるアノテーションや撮像方法の問題点を主張



手動アノテーションの問題点



デバイス装着による不自然さ

GFIE [Zhengxi+, CVPR2023]

データセット構成の観点

従来データセットにおけるデータ数や多様性の不足を主張

Dataset	# Sites	# Species	# Videos
KARB [26]	1	3	742
WildLive [16]	1	3	22
CCR [4]	3	1	27
SA-FARI(ours)	741	99	11,609

SA-FARI [Dante+, CVPR2026]

データセット構築中：再現性の情報はどこに・どれだけ書くべきか

※CVPR2026のタイトルに“Dataset”を含む論文の内の72本を対象

■ 本文のみ ■ 本文+Supp ■ Suppのみ ■ 記載なし

本文・Suppに含まれる割合

データセット収集方法

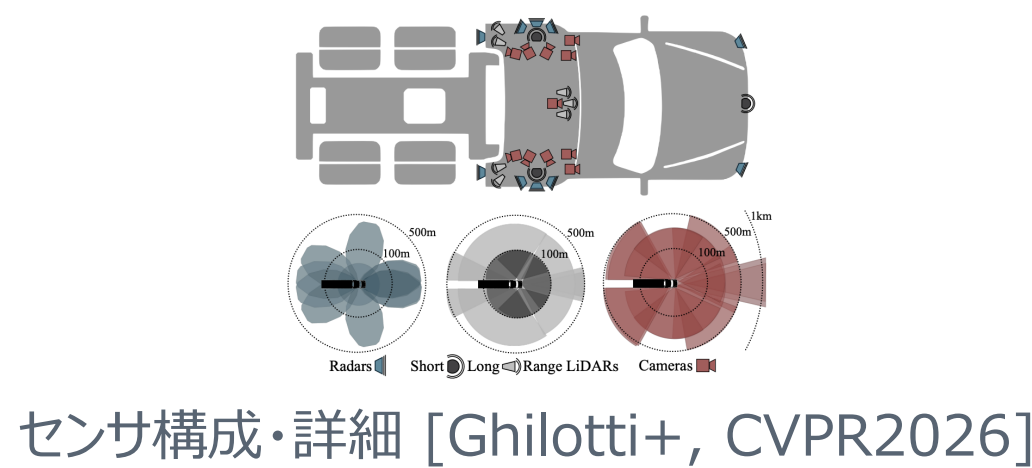
撮影型

- 撮影環境設定
- センサ構成・詳細
- 収集品質評価

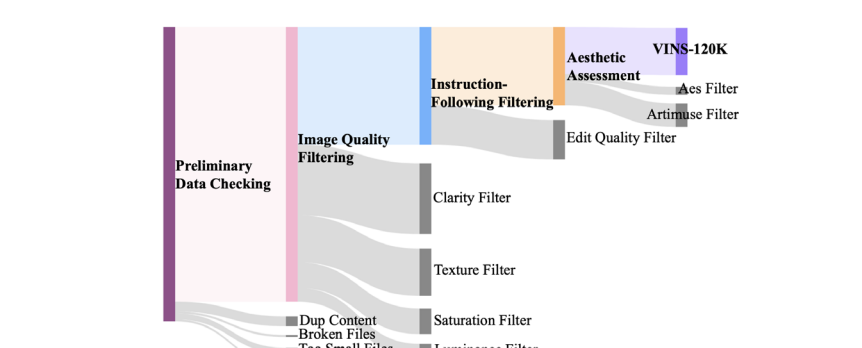
Web・既存データ活用型

- 収集ソース
- フィルタ条件・方法
- 実装の詳細

本文・Suppに含まれる割合



センサ構成・詳細 [Ghilotti+, CVPR2026]



フィルタリング [Chen+, CVPR2026]

本文には要点、Suppに詳細設定 が定石

アノテーション方法

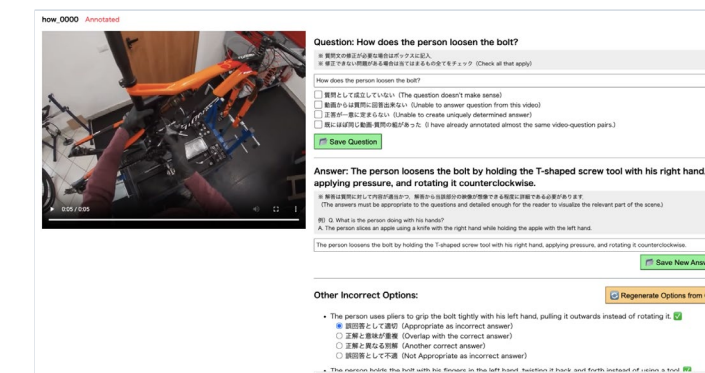
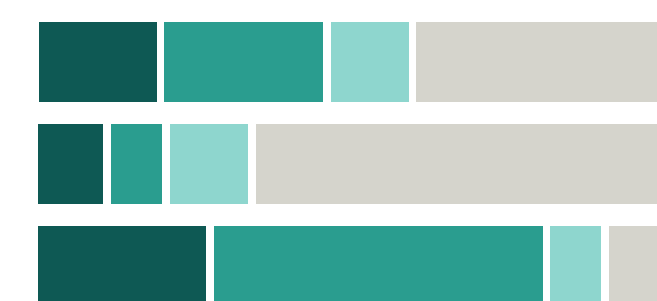
人手アノテーション

- アノテーターの属性
- ツール・システム
- 品質評価

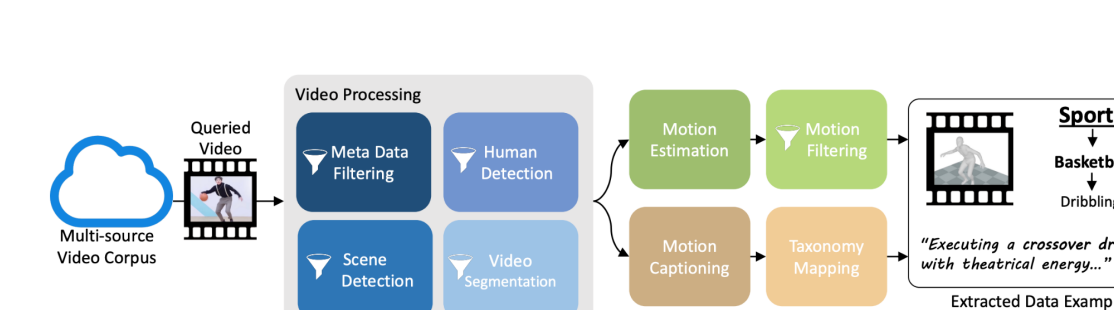
自動アノテーション

- パイプライン図
- 使用モデル
- モデルの詳細設定

本文・Suppに含まれる割合



アノテーションUI [Tateno+, CVPR2026]



パイプライン [Zhang+, CVPR2026]

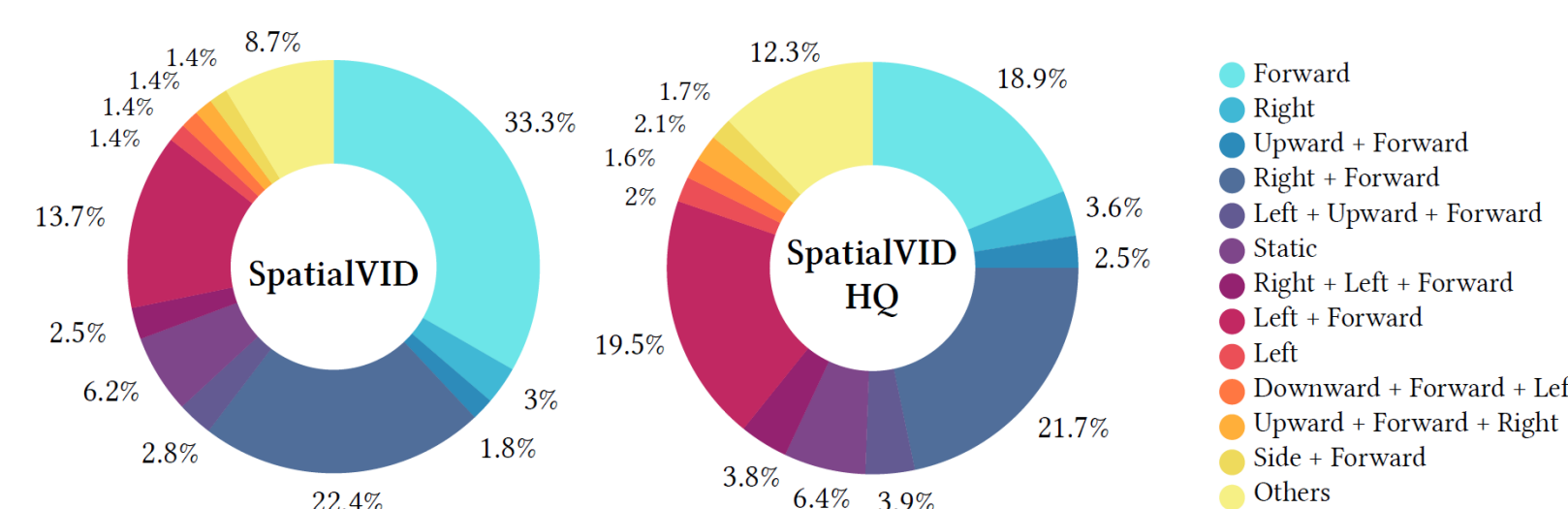
ツールや品質評価を記載することで差別化

データセット構築後：構成と品質を分析し利用へつなげる

データセットの分析

データセットの構成

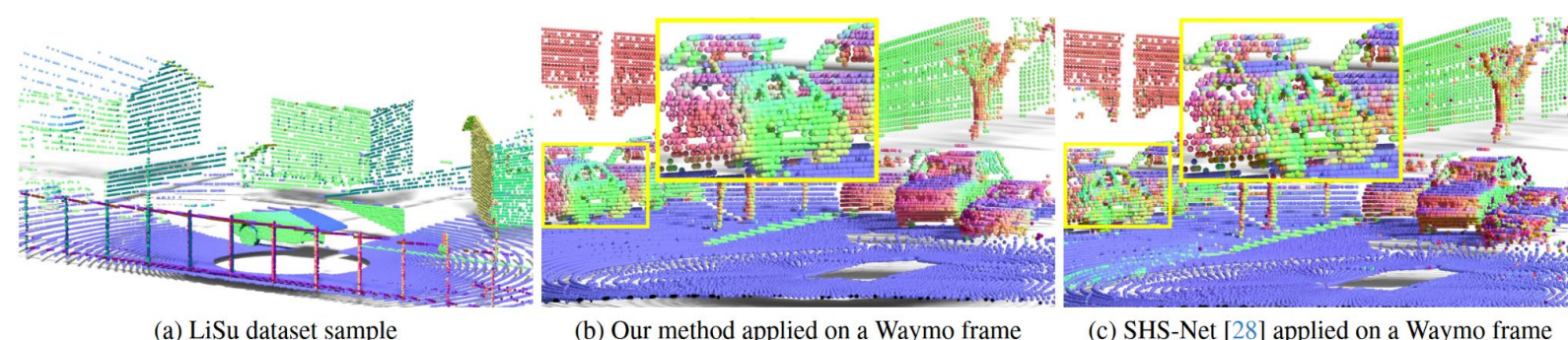
- 具体的なデータ例
- データ分布・多様性
- メタデータの詳細



データ分布・多様性 [Wang+, CVPR2026]

既存データセットとの比較

- データ形式・規模・対応タスク
- 定性的な品質比較
- ベンチマーク実験の比較



定性的な品質比較 [Dušan+, CVPR2025]

データセットの公開

見つけやすさ

- プロジェクトページの公開
- 動画やビューアでの可視化
- arXivへの登録

使いやすさ

- フォルダ構成の明示
- 導入手順の詳細な説明
- サンプルコードの充実

OLATverse: A Large-scale Real-world Object Dataset with Precise Lighting Control

CVPR 2026 (Oral, Best Paper Award Candidate)

Xilong Zhou¹, Junchen Chen¹, Pramod Rao¹, Timo Teufel¹, Linjie Lyu¹, Tigran Mnatsian¹, Oleksandr Sotnychenko¹, Xiao-Xiao Long², Marc Habermann¹, Christian Theobalt¹

¹Max Planck Institute for Informatics, ²Nanjing University
(^{*} Equal contribution)

arXiv Supplementary Material Code Preview Data Full Data Video



プロジェクトページ [Xilong+, CVPR2026]