どちらの「トーカー」が優れているか？

AI生成トーキングヘッドの主観的・客観的品質評価

Yingjie Zhou, Jiezhang Cao, et al.
Shanghai Jiao Tong University, PengCheng Laboratory, Harvard Medical School

NotebookLM

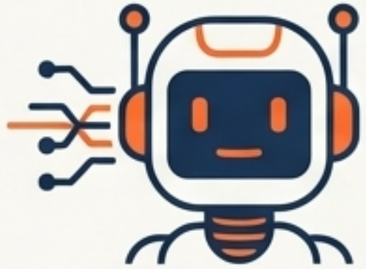
デジタルヒューマン制作のパラダイムシフト

従来の手法

手作業によるモデリングとアニメーションは、膨大な時間と専門知識を要し、コストが高く、スケーラビリティに欠ける。

時間がかかる | 高コスト | 専門的

AIによるアプローチ



Text-to-Image (T2I)モデルと音声駆動の「トーカー」により、生成が高速化・自動化される。しかし、その品質は未知数である。

高速 | 自動化 | 品質？

NotebookLM

新たな課題：AI生成品質の「ものさし」が存在しない

- **主観に依存:**
  品質評価は個人の感覚に頼っており、一貫性がない。

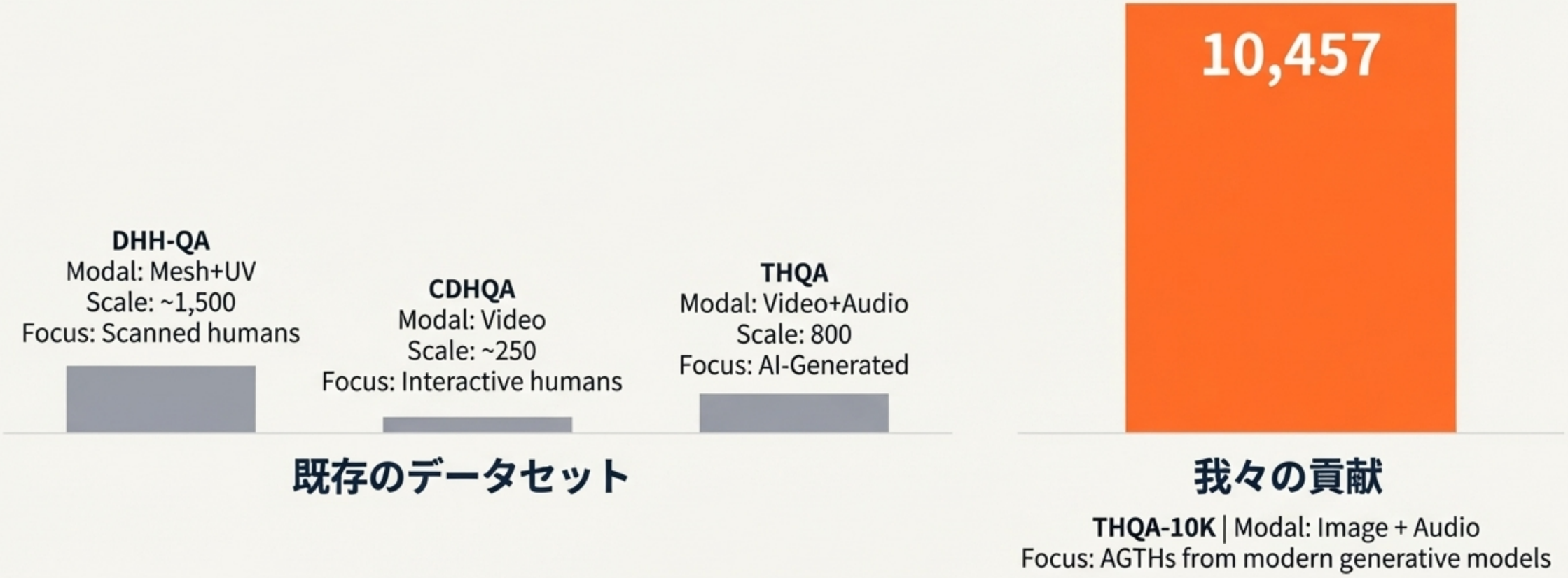
- **比較が困難:**
  異なる手法（Talker）間で、公正な性能比較ができない。

AIは多くの「トーカー」を生み出したが、どれが本当に優れているのか？

- **開発の停滞:** 明確な評価指標がなければ、技術の改善サイクルが回らない。

NotebookLM

既存のデータセットでは、
現代のAGTH評価ニーズに対応できない



規模、多様性、そして現代の生成モデルへの焦点において、決定的なギャップが存在した。

NotebookLM

貢献①：AGTH品質評価の決定版データセット「THQA-10K」を構築

10,457

AI-Generated Talking Heads

12

主要なText-to-Imageモデル

14

先進的な音声駆動「トーカー」

これまでで最大かつ最も包括的なAGTH品質評価データセット。

14のプロンプトと70の音声データから生成。

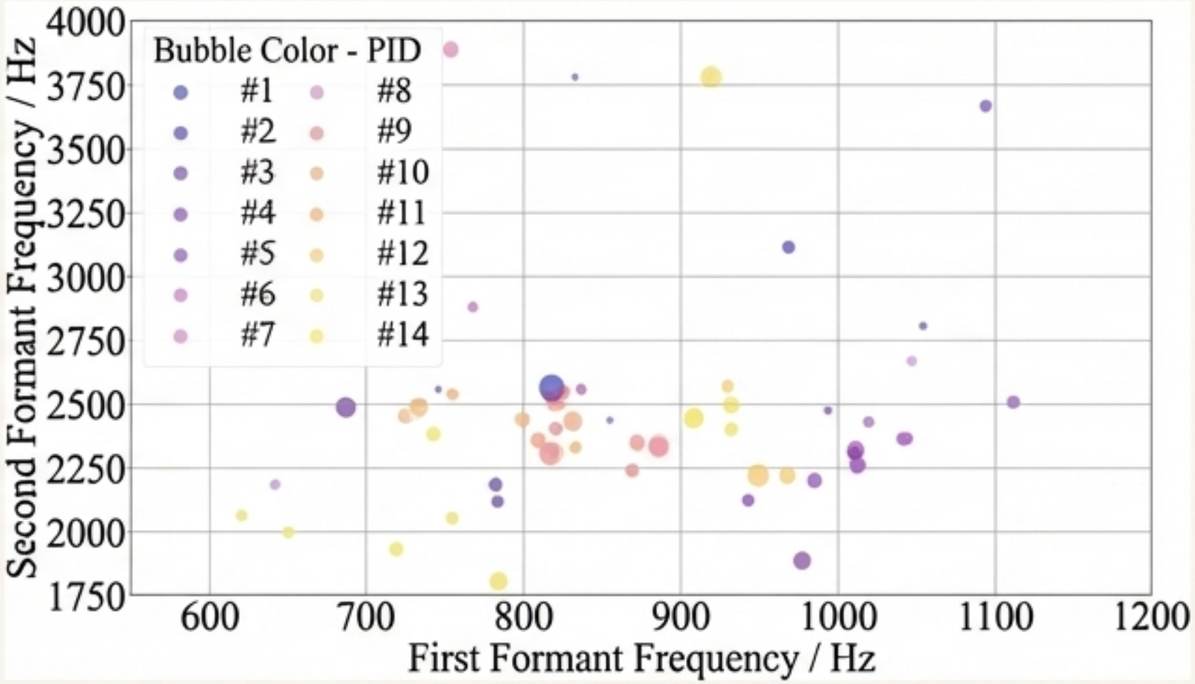
NotebookLM

多様性と代表性を確保したデータセット設計

多様な人物像を生成するプロンプト

年齢、性別、写真スタイルを網羅する14のプロンプト。高品質（8k）かつリアルな人物像の生成を目指す。

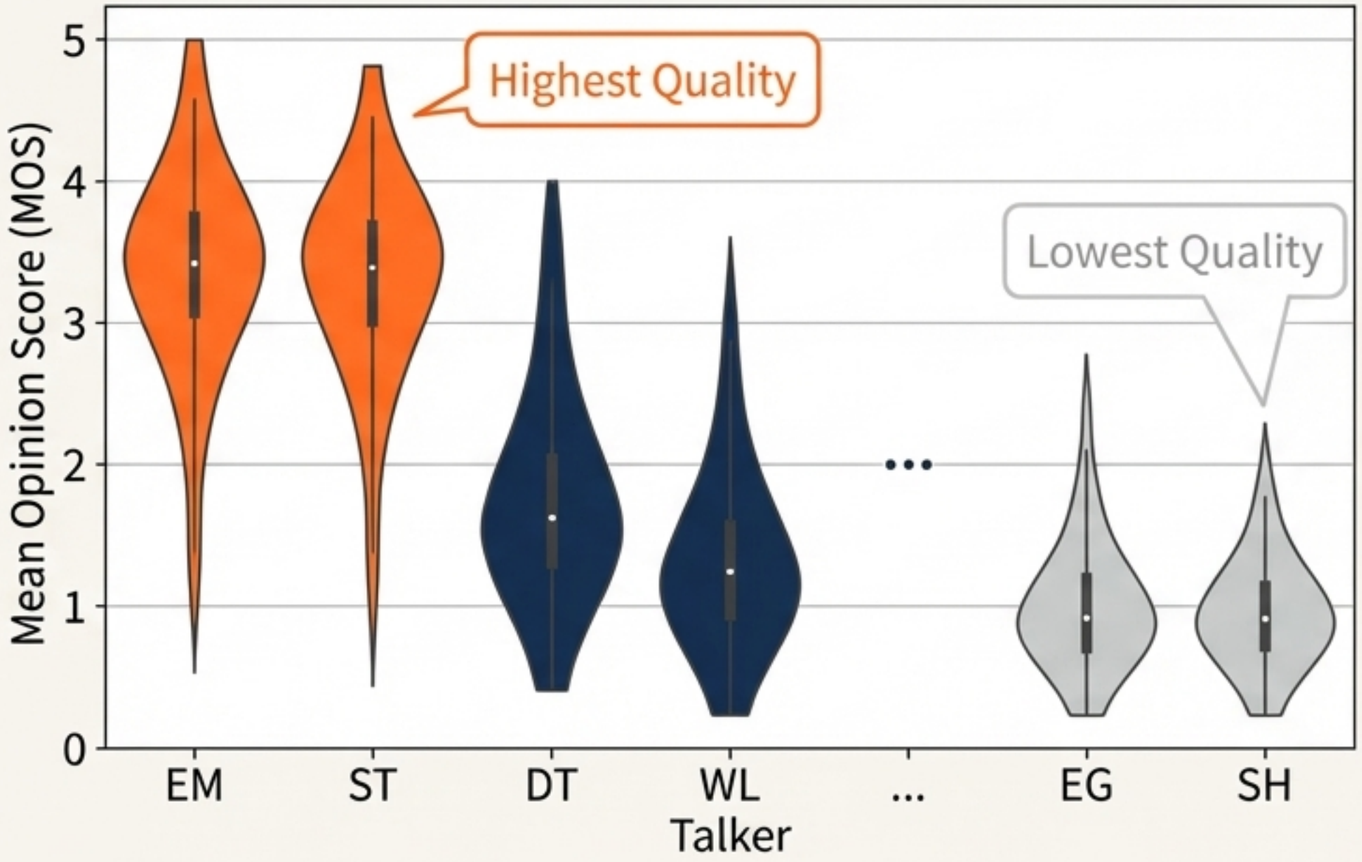
豊かな音素を含む音声データ



各プロンプトに対応する5つの音声（3〜10秒）。多様な口の形や舌の動きを反映する音響特性を持つ。

NotebookLM

データが示す真実：「トーカー」の選択が品質を決定づける



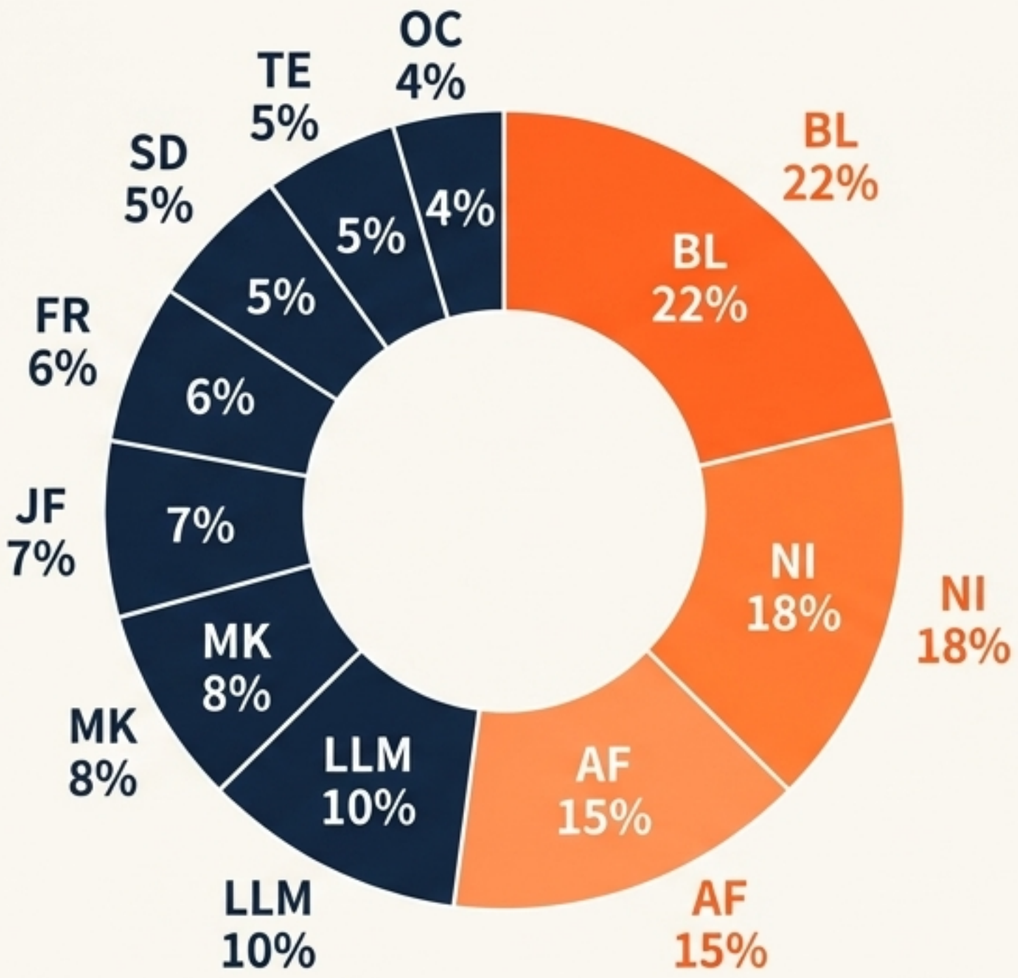
T2Iモデルやプロンプトの影響は限定的: 人物画像（T2I）やセリフ（Prompt）の違いによる品質差は比較的小さい。

「トーカー」の性能差が顕著: 使用する「トーカー」によって、生成されるAGTHの品質に大きなばらつきが生じる。

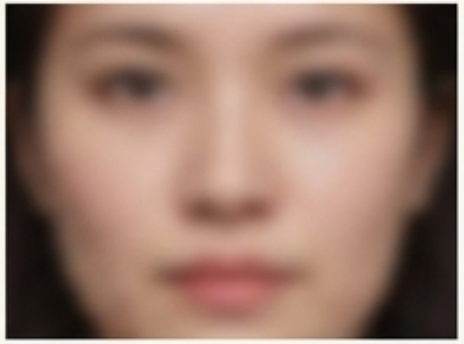
品質の鍵は、どの「顔」を使うかではなく、どの「話し手」を選ぶかにある。

NotebookLM

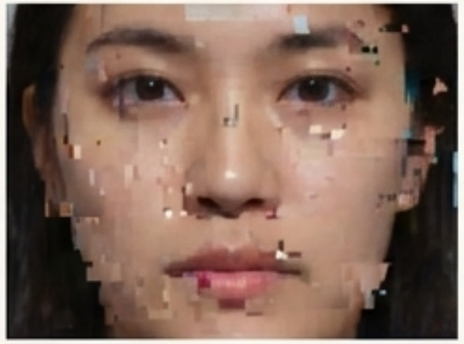
AGTHに潜む10種類の品質劣化（ディストーション）



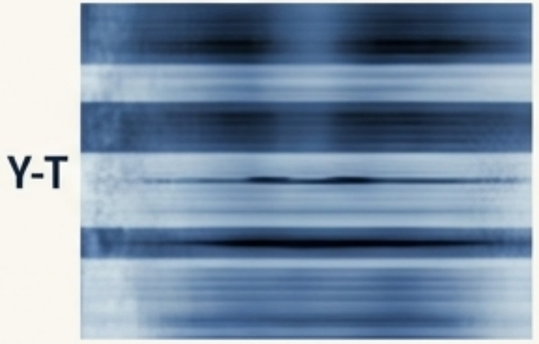
Blur (BL)



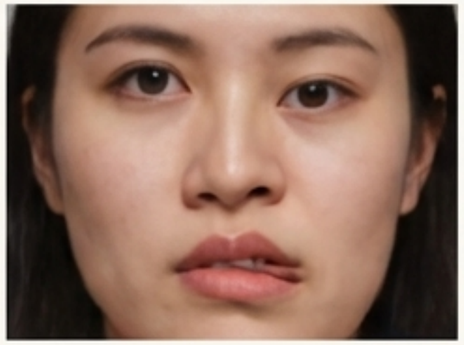
Artifacts (AF)



Little Lip Motion (LLM)



Misaligned Keypoints (MK)

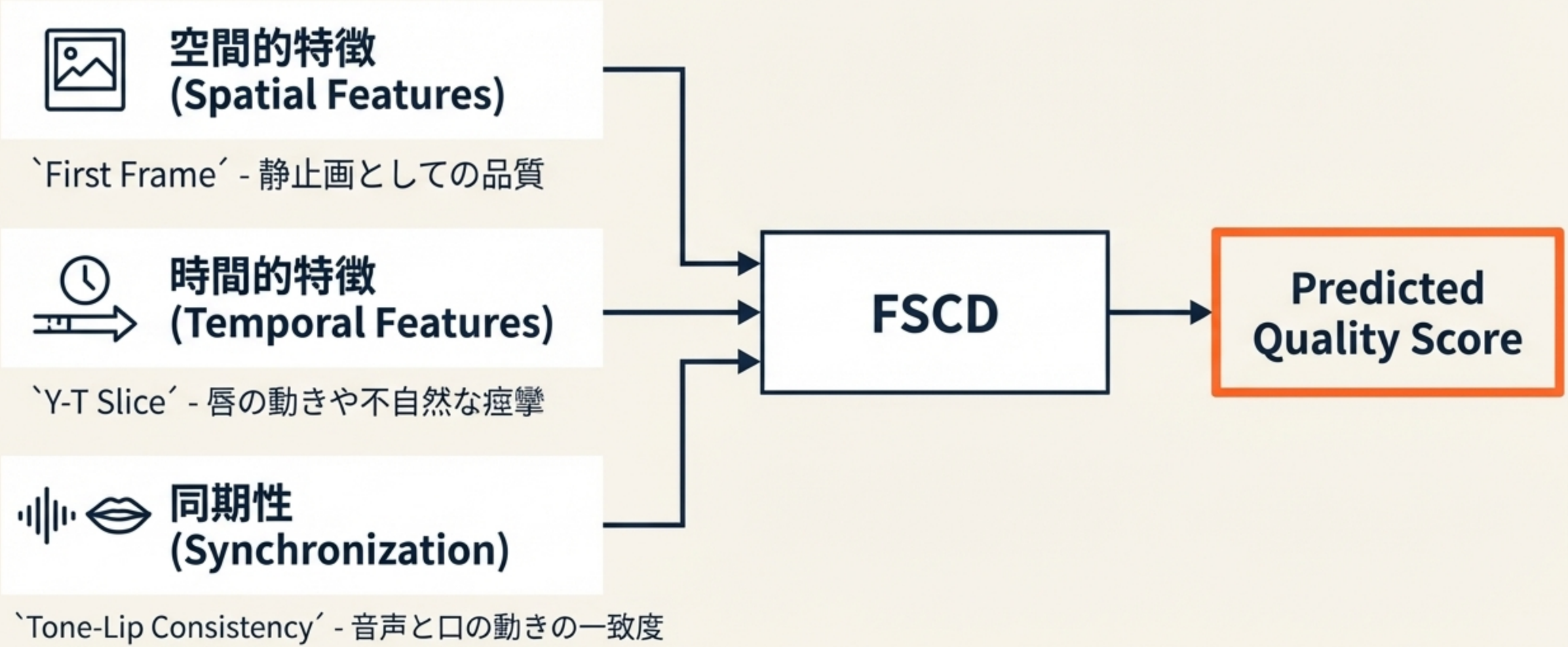


一つのAGTHが複数のディストーションを併発することが多い。
特に「キーポイントのずれ(MK)」は知覚品質への**影響が最も大きい**。

NotebookLM

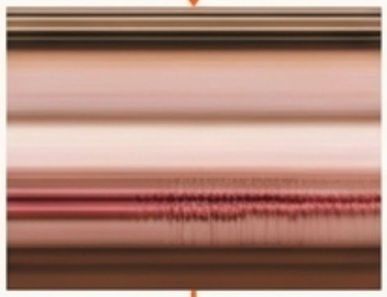
貢献②：客観的品質評価手法「FSCD」を提案

人間の知覚を模倣し、AGTHの品質を多角的に評価する。



NotebookLM

FSCDのフレームワーク：空間・時間・同期性の統合



1. First Frame (空間)

最初のフレームから静的な画質を評価。**Swin Transformer**が特徴を抽出。

2. Y-T Slice (時間)

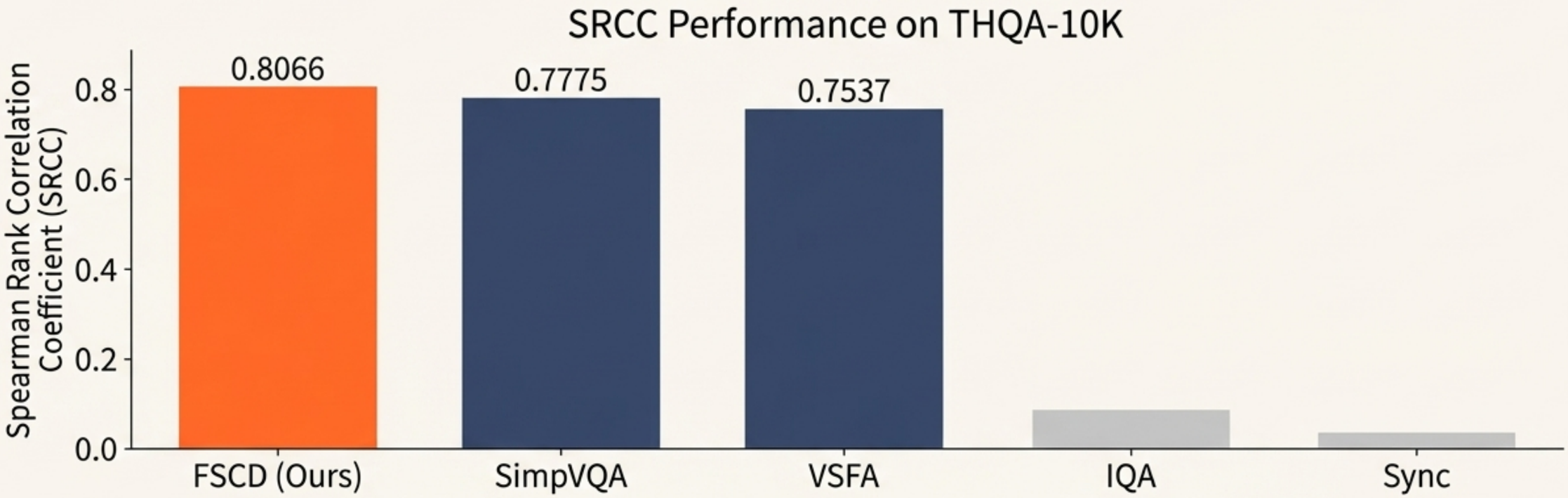
顔の中心を通る縦のスライスを時系列で並べ、動きのパターンを画像化。**LLM**や**MT**といった時間的な歪みを効率的に捉える。

3. Tone-Lip Consistency (同期)

SyncNetを用いて、音声のMFCCと口の領域の動きを比較。リップシンクのズレ（**LSE-C**, **LSE-D**）を定量化。

NotebookLM

性能比較：FSCDは既存手法を全てのデータセットで凌駕する

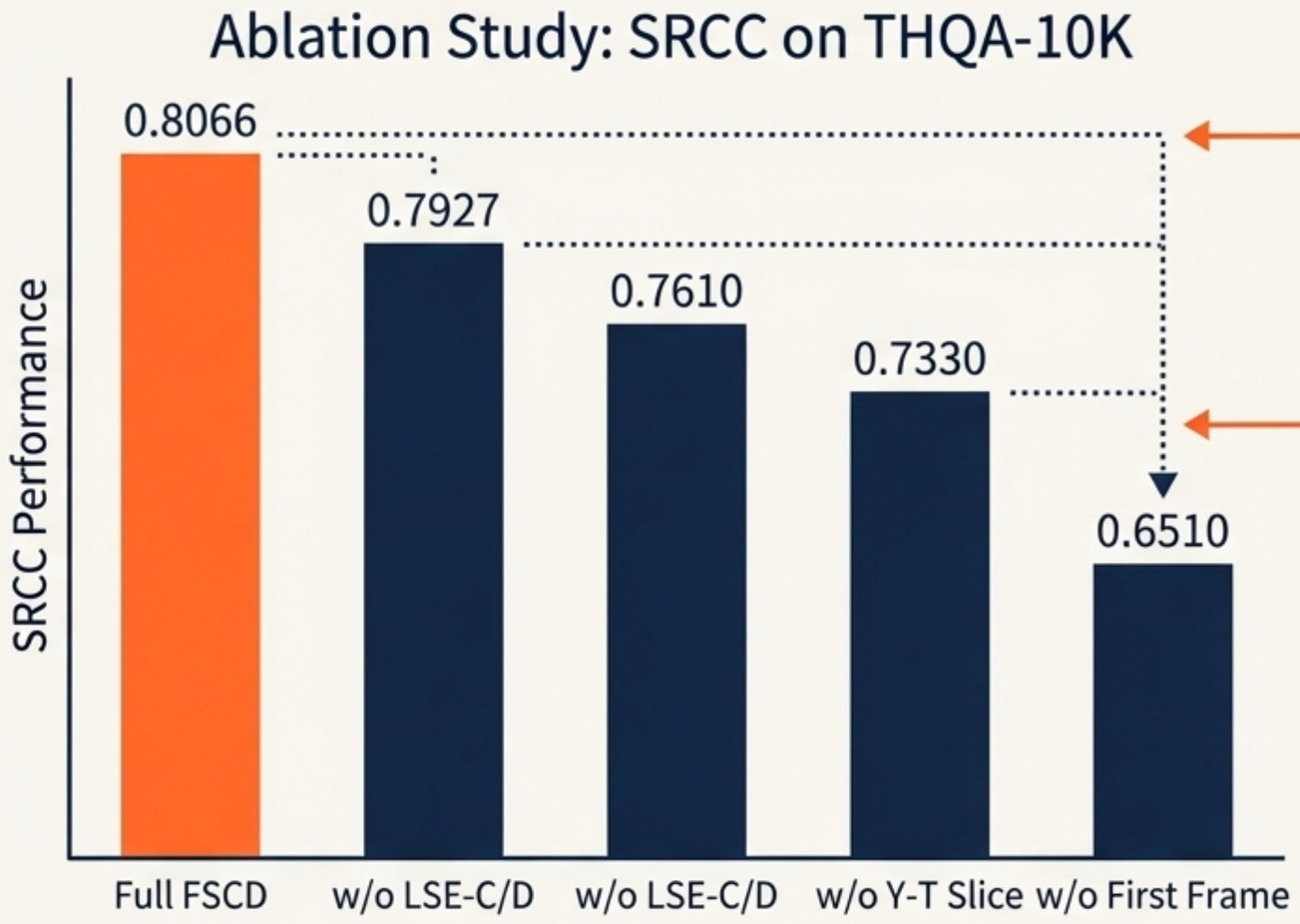


Method	SRCC↑	PLCC↑	KRCC↑	RMSE↓
FSCD (Ours)	**0.8066**	**0.8322**	**0.6228**	**0.6333**
SimpVQA	0.7775	0.8039	0.5931	0.6832
VSFA	0.7537	0.7754	0.5726	0.7343

FSCDは、人間の主観評価と最も高い相関を示し、信頼性の高い客観的指標であることを証明した。

NotebookLM

なぜFSCDは優れているのか？各構成要素の重要性を検証



空間的特徴 (First Frame)が最も重要
静止画の品質が、全体の知覚品質の基盤となっている。

時間的特徴 (Y-T Slice)も不可欠
動きの自然さを評価するために、Y-T Sliceが効果的に機能している。

全ての要素が性能向上に寄与
空間、時間、同期性の組み合わせが、相乗効果を生み出している。

NotebookLM

未来へ：AGTH研究を加速させるための羅針盤

THQA-10K Dataset

貢献：10,457件のAGTHを含む、大規模かつ包括的なベンチマークを確立。

「トーカー」開発のための新たな「ゴールドスタンダード」を提供。公正で再現性のある比較を可能にする。

FSCD Method

貢献空間、時間、同期性を統合した、SOTAの客観的品質評価手法を開発。

影響：研究者が開発サイクルを高速化するための信頼できるツールを提供。手動評価のコストと時間を削減する。

本研究は、次世代の高品質なデジタルヒューマンの実現に向けた、客観的で信頼性の高い基盤を構築するものである。

Dataset and code available at: https://github.com/zyj-2000/Talker

NotebookLM