

文化圏を考慮した衣服画像生成のための データ駆動型プロンプト自動チューニング

キヤノン株式会社

井関茜

グローバルな画像生成VLMの課題

各文化の言葉のニュアンスに合わせ画像生成したい。
「日本人にとって高級感」のような単純なプロンプト調整では、下記のような画像が生成されることがある。

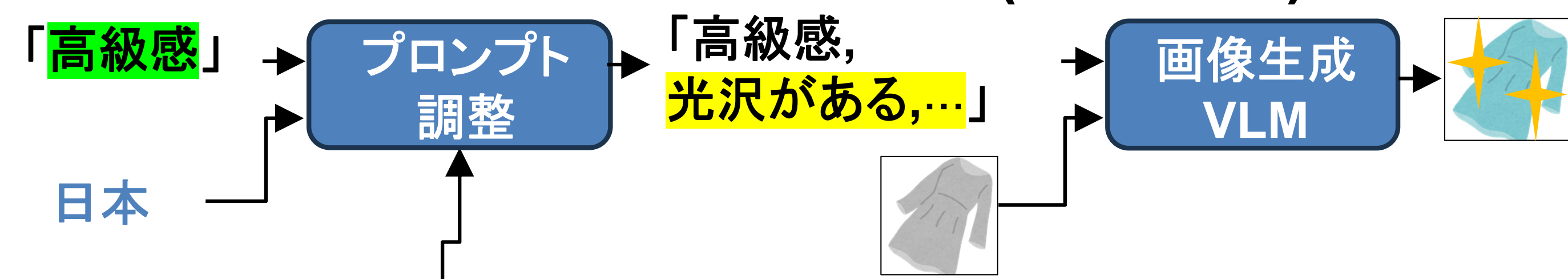
- 世界平均的な高級感
- 他国から見たステレオタイプ（桜など）
- 高級感にもいろいろあるのに、ワンパターン

手法・ツールの適用による解決

ウェブサイトから収集したデータに基づいて、
各文化特有の言葉のニュアンスを、文化による意味の違いが少ない画像特性にかみ砕いた表現に変換するモデルを構築し、プロンプト調整を行う。
それにより、画像生成VLM自体にバイアスがあっても、意図通りの画像生成を行う。

アプローチ

各文化の抽象表現（画像の印象を表す言葉）を、文化による意味の違いが少ない具体表現（画像特性）に変換



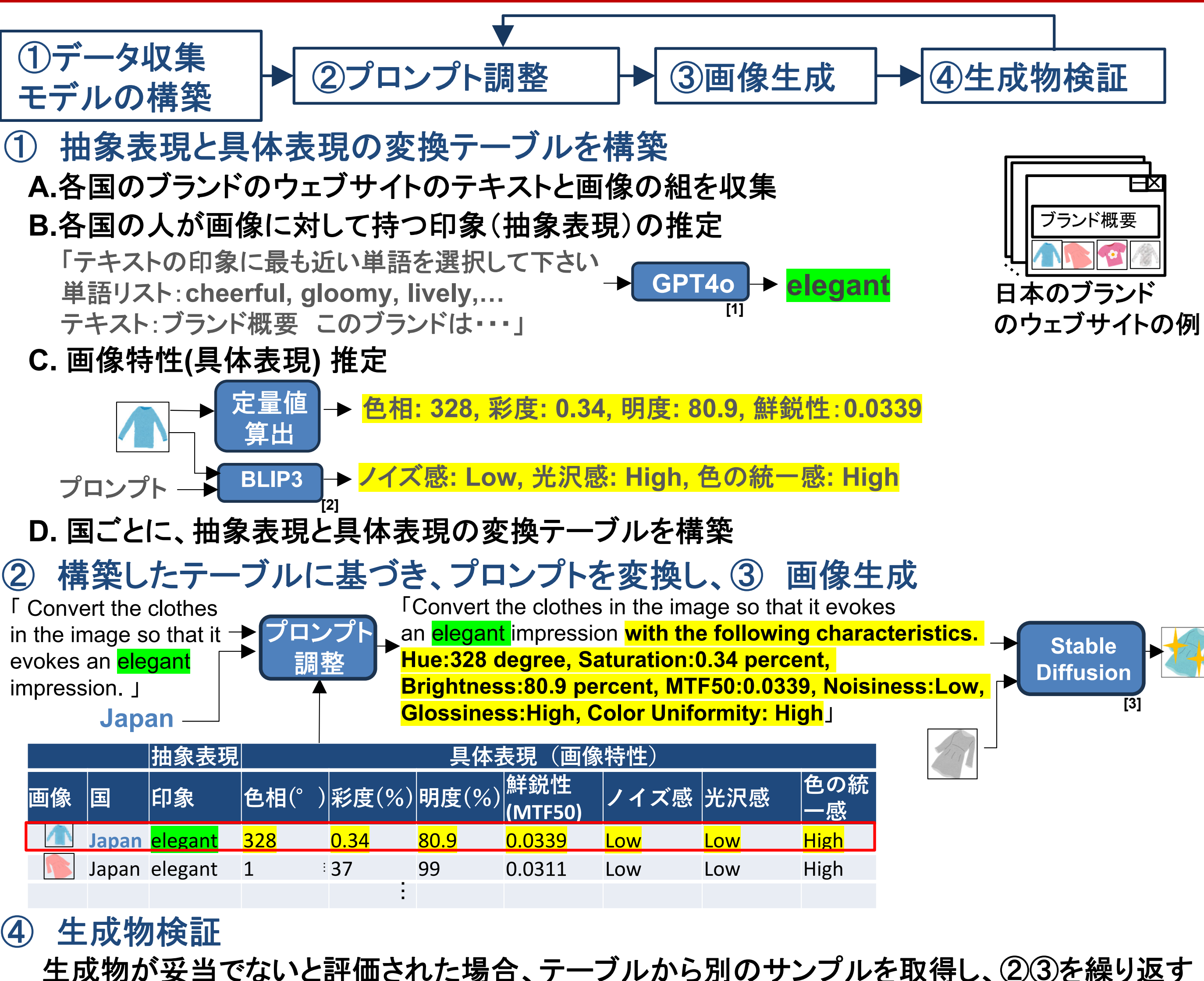
国	抽象表現		具体表現（画像特性）						
	印象		色相	彩度	明度	鮮鋭性	ノイズ感	光沢感	色の統一感
日本	高級感		15	20	67	0.0339	Low	High	High
日本	高級感		337	50	91	0.0246	Low	Low	High
タイ	高級感		0	0	83	0.0135	Low	Low	High

各文化の人へアンケートせず、各国のブランドのウェブサイトのテキストと画像から変換モデルを低コストに構築

<仮説1>ある国の高級ブランドの画像はその国の人にとって高級な印象を持つはず

<仮説2>各ブランドが与えようとしている、高級感などの印象は、ウェブサイト上のテキストから推定できるはず

提案システム全体像



評価結果

- 各国の人向けにelegantな画像を生成し、評価した。提案システムは、単純なプロンプトと比べ、ステレオタイプを軽減し、多様な画像を生成できた
- 機械的な評価: 単純なプロンプトと比べ、生成画像の画像特性が各国のelegantな画像の分布に収まった

	日本	タイ	中国	ベトナム
単純なプロンプト 例) Convert the clothes in the image to evoke an elegant impression in Japan.				
提案手法				

- 人間による評価: 6パターンの入力画像に対する日本人向け生成画像を、日本人24名に比較させた提案システムは、単純なプロンプトと比べ、elegantさは維持しつつステレオタイプを軽減し、かつ、より商品として魅力的な画像を生成できた

	評価項目	提案手法の勝率
単純なプロンプトと比べて	日本らしいステレオタイプを軽減	83.3% (5/6)
	elegantという言葉の印象に近い	50% (3/6)
	商品として魅力的	83.3% (5/6)

結論

- ウェブサイトのテキストと画像から、各国の文化の感性を抽出できる可能性を示した
- 各文化の感性に即した多様な画像生成が可能になった
- ステレオタイプを軽減し商品企画等への応用性が向上

展望

- 因果推論の活用: ある顧客をターゲットとした商品開発で、生成画像が選ばれた理由を説明可能に
- 評価拡充、各文化のペルソナ作成と評価自動化
- Human-in-the-Loopによるモデルの更新
- 画像生成VLMのコントロール性向上

[1] Achiam et al. Gpt-4 technical report. arXiv:2303.08774, 2023

[2] Xue et al. BLIP-3: A Family of Open Large Multimodal Models. arXiv:2408.08872, 2024

[3] Rombach et al. High-resolution image synthesis with latent diffusion models, CVPR2022