Automatic Aesthetic Evaluation in Generative Image Models

Larissa D. Gomide, Lucas N. Ferreira, Wagner Meira Jr.

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

ICCC'25

PROGRAMA DE PÓS-GRADUAÇÃO
CIÊNCIA DA COMPUTAÇÃO
UFMG

U F M G
UNIVERSIDADE FEDERAL
DE MINAS GERAIS
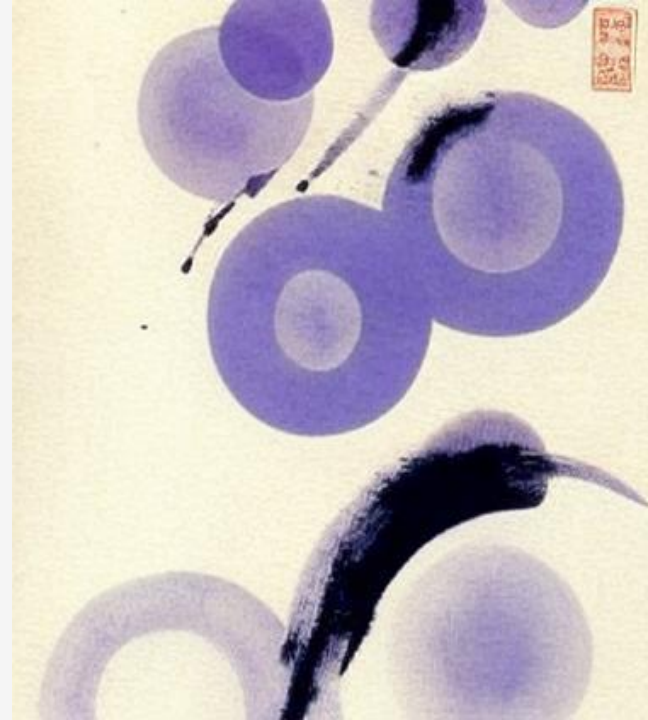
Background

- **APDDv2 Dataset**
  - ~10K human visual artwork
  - 10 aesthetic attributes
    - Ex: Light and Shadow
  - Expert labelling (artists)

- **ArtClip model** (Fine-tuned CLIP)
  - Evaluated only on human-created pieces

[Jin et al., 2024]
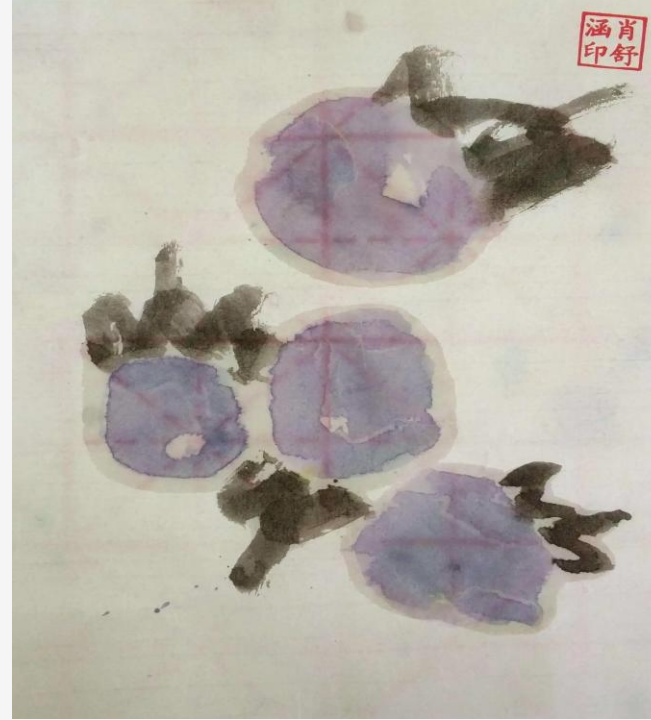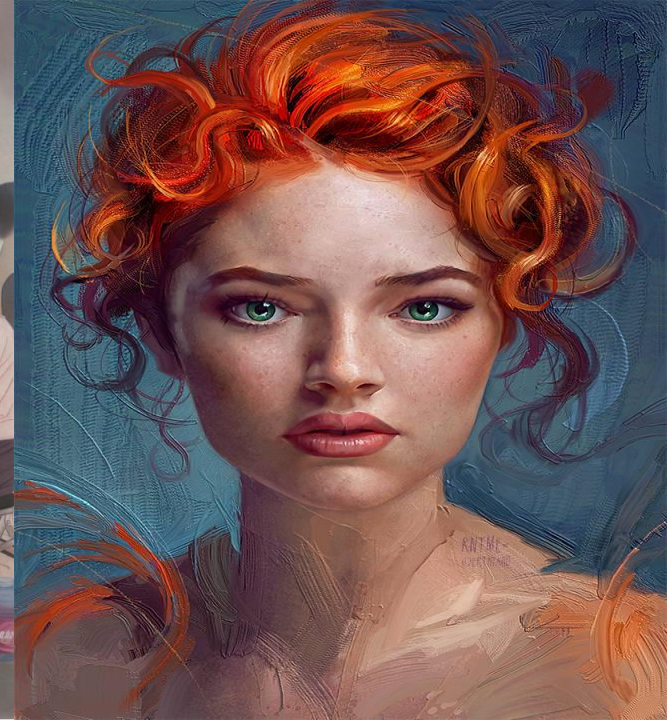
Main Contribution

Evaluate ArtClip as an **objective metric** for AI-generated artwork

[Gomide, N. Ferreira, and Meira Jr. et al., 2025]

APDDv2

	Artistic Categories	Theme & logic	Creativity	Layout & Composition	Space & Perspective	Sense of Order	Light & Shadow	Color	Details & Texture	The overall	Mood
	oil painting, romanticism, portraiture	3.86	4.14	4.14	-	3.86	-	4.14	3.86	4.14	-
	traditional Chinese painting, freehand,floral and avian	6	5.83	6	-	5.83	-	6.33	5.33	5.83	6.33
	sketching, symbolism, portraiture	-	-	7.67	7.83	-	8	-	7.5	7.83	-
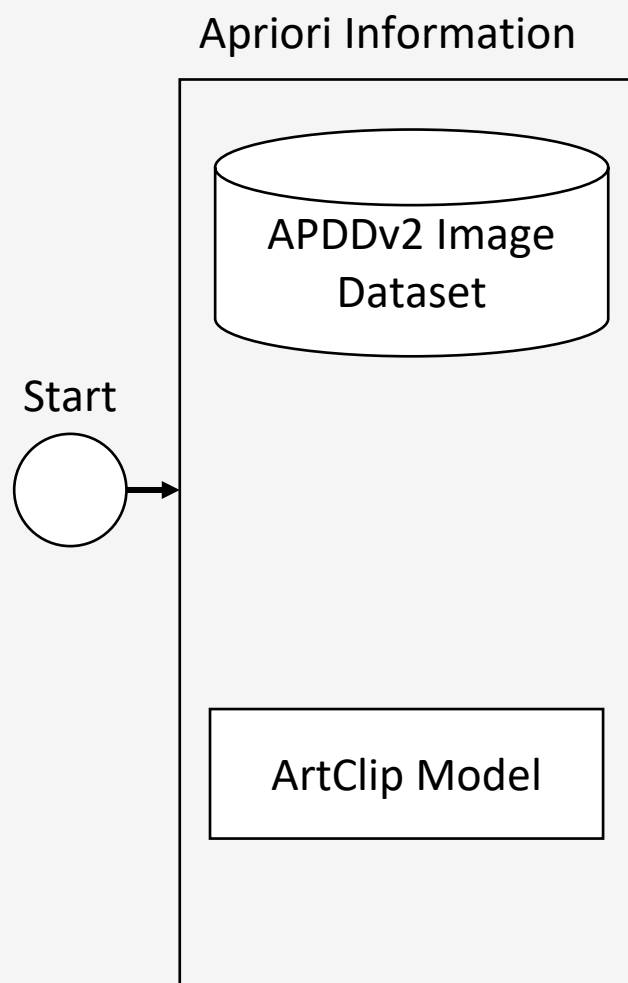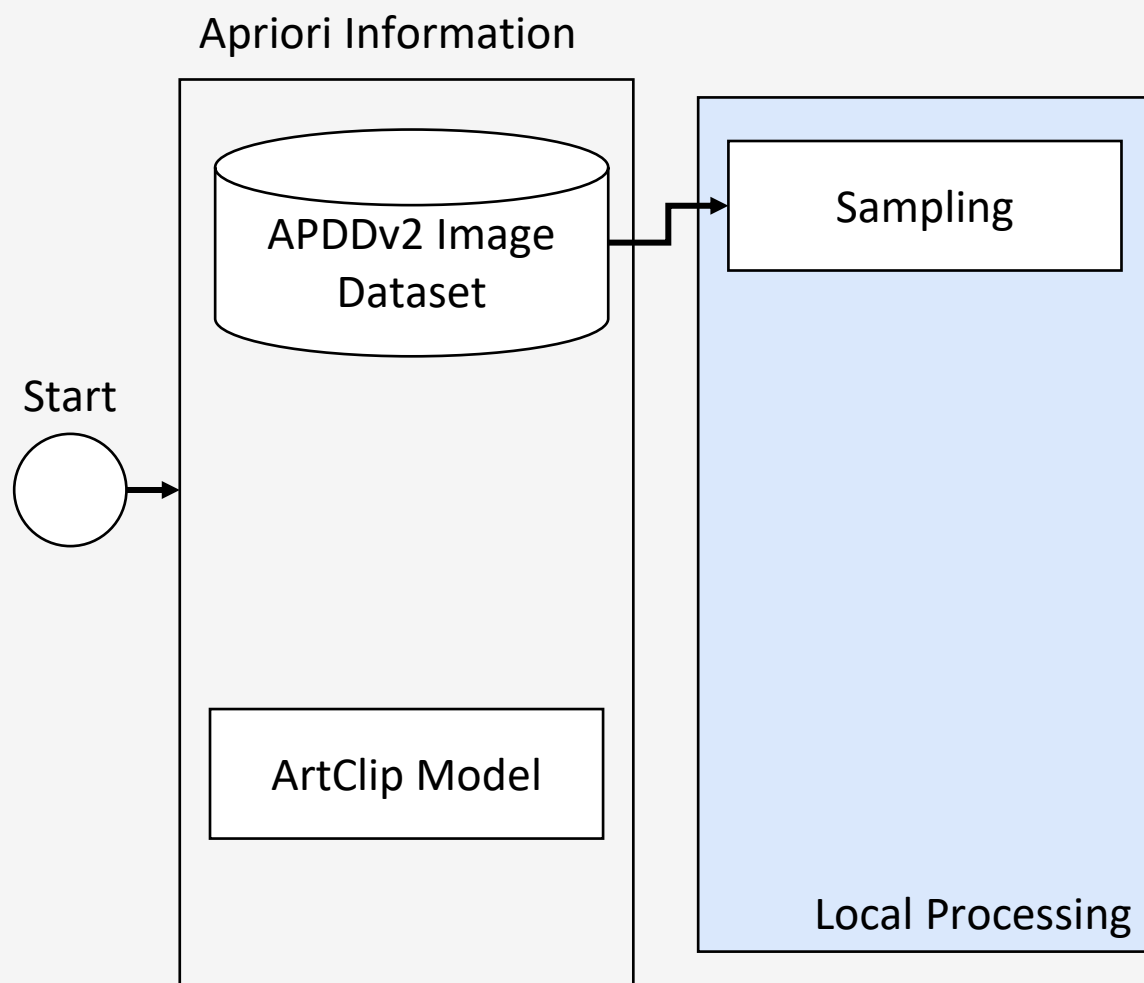
Before
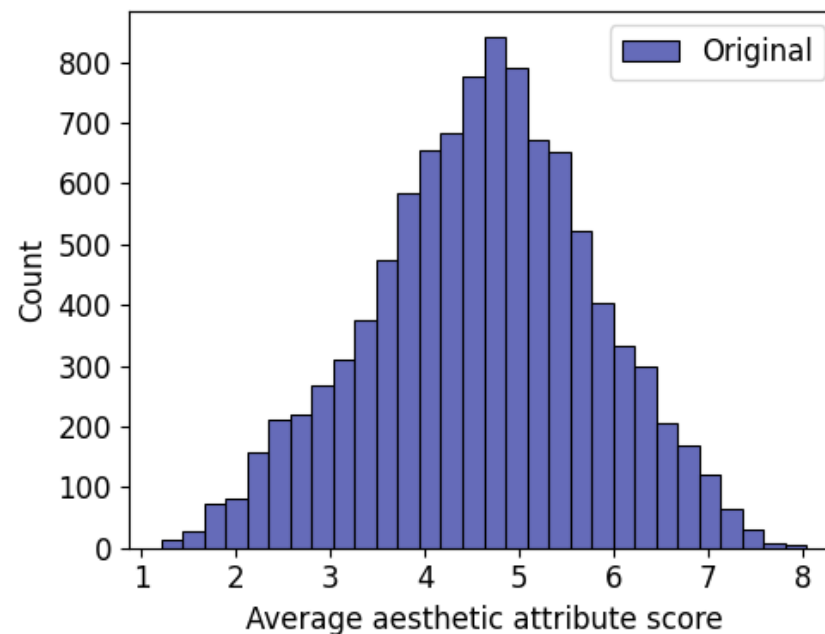
After

10,023 paintings
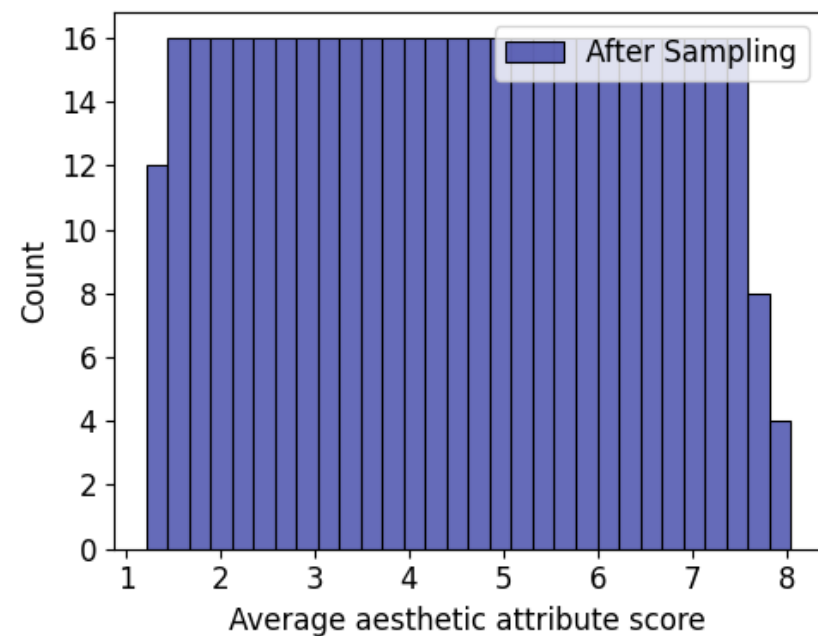
~500 paintings
Statistical Relevance

Image-to-text
deepseek

Text-to-image
deepseek

Scoring

GPU
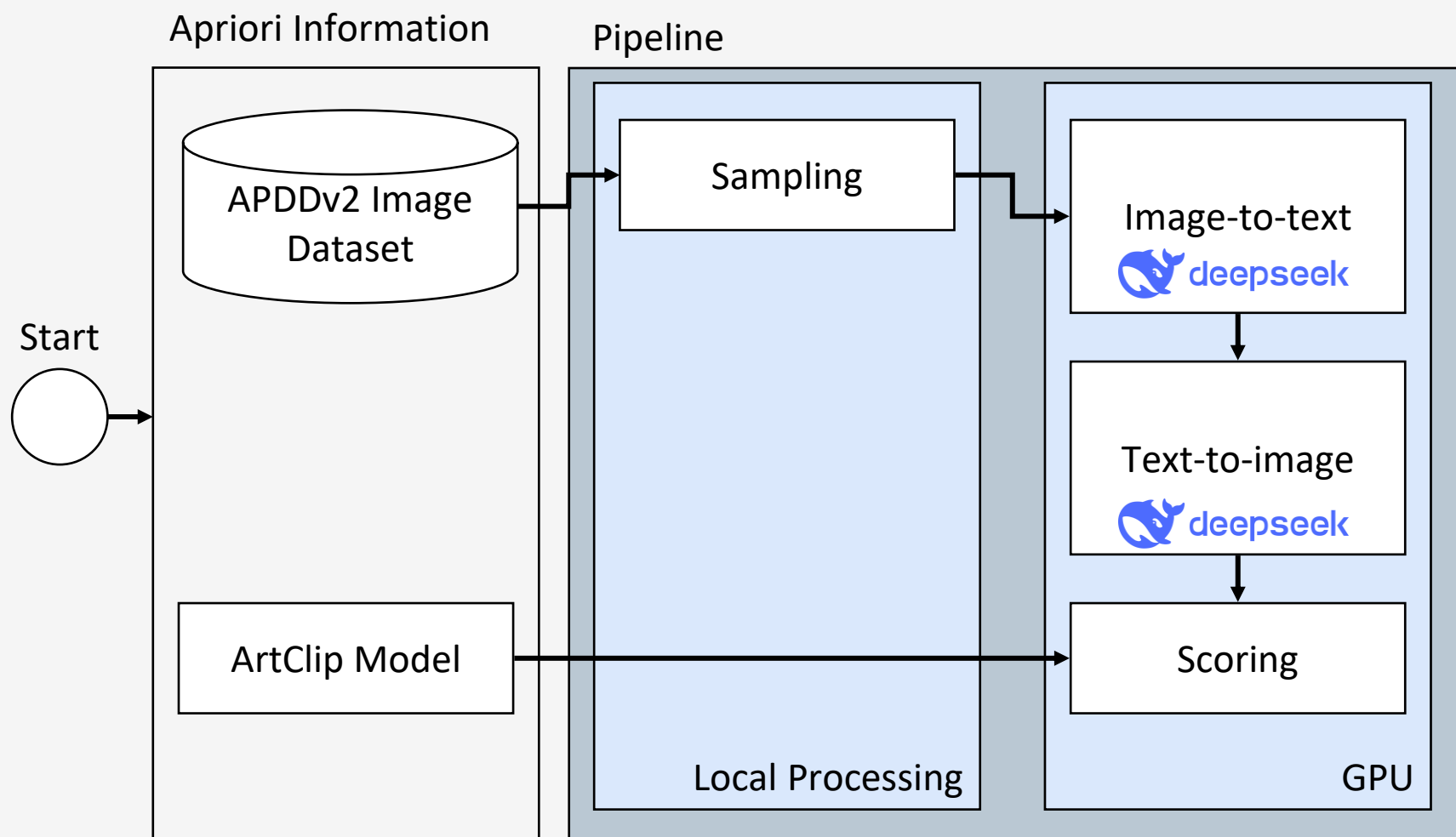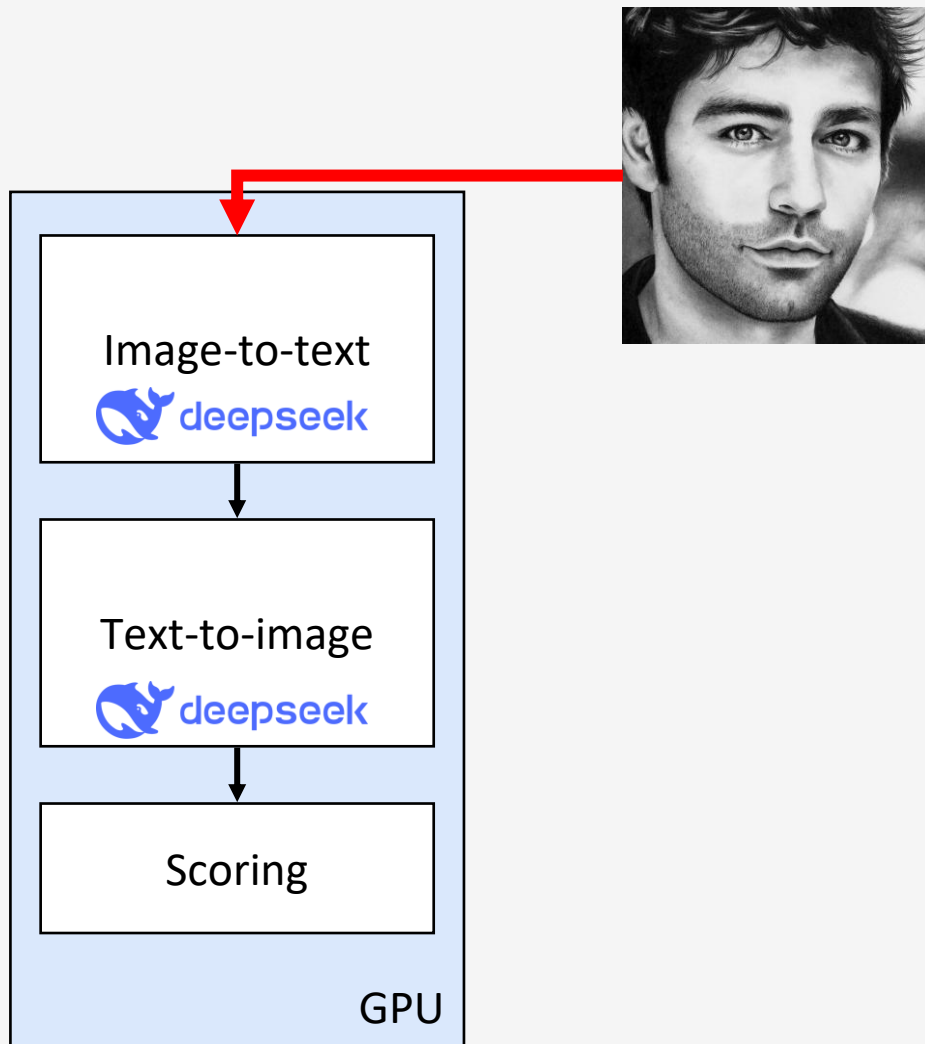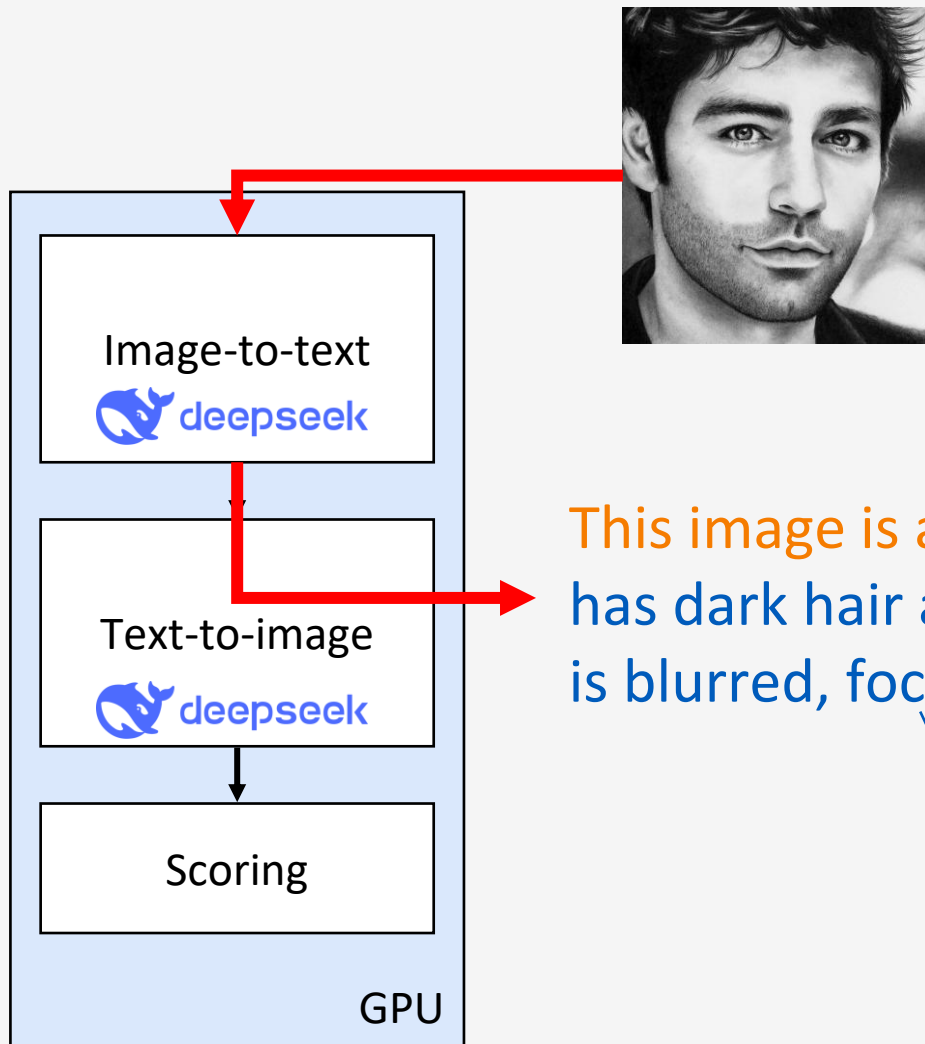
Style

This image is a black and white portrait of a person. The individual has dark hair and is looking directly at the camera. The background is blurred, focusing attention on the person's face.

Content

This image is a black and white portrait of a person. The individual has dark hair and is looking directly at the camera. The background is blurred, focusing attention on the person's face.

10 Aesthetic Attributes of APDDv2*

Human

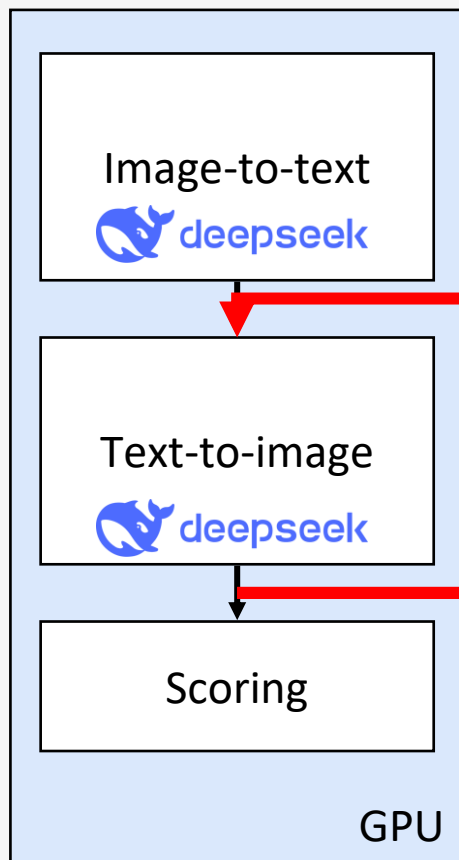|	Artistic Categories	Total aesthetic score	Theme and logic	Creativity	...	Mood
	sketching, symbolism, portraiture	80.83	-	-	...	-

Janus-1B

	Artistic Categories	Total aesthetic score	Theme and logic	Creativity	...	Mood
	sketching, symbolism, portraiture	75.06	6.09	5.41	...	5.83

Janus-7B

	Artistic Categories	Total aesthetic score	Theme and logic	Creativity	...	Mood
	sketching, symbolism, portraiture	82.29	6.35	6.21	...	5.95

10 Aesthetic Attributes of APDDv2*

Human

Artistic Categories	Total aesthetic score	Theme and logic	Creativity	...	Mood
sketc... symb... portr...	82.83	-	-	...	-

Janus-1B

Artistic Categories					Mood
sk... s... p...					5.83

Janus-7B

Artistic Categories	Total aesthetic score	Theme and logic	Creativity	...	Mood
sketching, symbolism, portraiture	82.29	6.35	6.21	...	5.95

Contribution: Paired dataset:
1. Human artwork
2. Janus (AI-generated) small
3. Janus (AI-generated) large

* Vide: Table 2 from the original article

13

Results
Model Size

- **Objective:** Analysis of the impact of model size on aesthetic score (relative to human artwork)

Attr.	Human	Janus-1B	Janus-7B
Theme and Logic (T&L)			
Creativity (Cre)			
Layout and Composition (L&C)			
Space and Perspective (S&P)			
Light and Shadow (L&S)			
Color (Col)			
Details and Texture (D&T)			
The Overall (TO)			
Mood (M)			

Results
Model Size

- **Objective:** Analysis of the impact of model size on aesthetic score (relative to human artwork)

- **Hypothesis:** A reliable aesthetic evaluator should assign **higher scores** to **human-created artworks**.

Attr.	Human	Janus-1B	Janus-7B
Theme and Logic (T&L)			
Creativity (Cre)			
Layout and Composition (L&C)			
Space and Perspective (S&P)			
Light and Shadow (L&S)			
Color (Col)			
Details and Texture (D&T)			
The Overall (TO)			
Mood (M)			

Results
Model Size

- **Objective:** Analysis of the impact of model size on aesthetic score (relative to human artwork)

- **Hypothesis:** A reliable aesthetic evaluator should assign **higher scores** to **human-created artworks**.

Attr.	Human	Janus-1B	Janus-7B
Theme and Logic (T&L)		$5.89^b \pm 0.86$	
Creativity (Cre)		$5.65^b \pm 0.81$	
Layout and Composition (L&C)		$5.69^b \pm 1.06$	
Space and Perspective (S&P)		$5.74^b \pm 1.06$	
Light and Shadow (L&S)		$5.53^b \pm 1.23$	
Color (Col)		$6.04^b \pm 0.98$	
Details and Texture (D&T)		$5.33^b \pm 1.25$	
The Overall (TO)		$5.76^b \pm 1.06$	
Mood (M)		$5.53^b \pm 1.00$	

Results
Model Size

- **Objective:** Analysis of the impact of model size on aesthetic score (relative to human artwork)

- **Hypothesis:** A reliable aesthetic evaluator should assign **higher scores** to **human-created artworks**.

| Attr.	Human	Janus-1B	Janus-7B
Theme and Logic (T&L)		$5.89^b \pm 0.86$	$6.14^c \pm 0.82$
Creativity (Cre)		$5.65^b \pm 0.81$	$5.82^c \pm 0.82$
Layout and Composition (L&C)		$5.69^b \pm 1.06$	$6.01^c \pm 1.01$
Space and Perspective (S&P)		$5.74^b \pm 1.06$	$6.09^c \pm 1.04$
Light and Shadow (L&S)		$5.53^b \pm 1.23$	$5.94^c \pm 1.21$
Color (Col)		$6.04^b \pm 0.98$	$6.20^c \pm 1.00$
Details and Texture (D&T)		$5.33^b \pm 1.25$	$5.72^c \pm 1.23$
The Overall (TO)		$5.76^b \pm 1.06$	$6.09^c \pm 1.02$
Mood (M)		$5.53^b \pm 1.00$	$5.85^c \pm 1.01$

Results
Model Size

- **Objective:** Analysis of the impact of model size on aesthetic score (relative to human artwork)

- **Hypothesis:** A reliable aesthetic evaluator should assign **higher scores** to **human-created artworks**.

- **Results: 1B < 7B < Human**

| Attr.	Human	Janus-1B	Janus-7B
Theme and Logic (T&L)	$6.25^a \pm 1.23$	$5.89^b \pm 0.86$	$6.14^c \pm 0.82$
Creativity (Cre)	$5.91^a \pm 1.41$	$5.65^b \pm 0.81$	$5.82^c \pm 0.82$
Layout and Composition (L&C)	$6.21^a \pm 1.35$	$5.69^b \pm 1.06$	$6.01^c \pm 1.01$
Space and Perspective (S&P)	$6.37^a \pm 1.38$	$5.74^b \pm 1.06$	$6.09^c \pm 1.04$
Light and Shadow (L&S)	$6.29^a \pm 1.45$	$5.53^b \pm 1.23$	$5.94^c \pm 1.21$
Color (Col)	$6.19^a \pm 1.38$	$6.04^b \pm 0.98$	$6.20^c \pm 1.00$
Details and Texture (D&T)	$6.10^a \pm 1.52$	$5.33^b \pm 1.25$	$5.72^c \pm 1.23$
The Overall (TO)	$6.26^a \pm 1.41$	$5.76^b \pm 1.06$	$6.09^c \pm 1.02$
Mood (M)	$6.03^a \pm 1.35$	$5.53^b \pm 1.00$	$5.85^c \pm 1.01$

Results
Consistency

- Computed the **average absolute difference** between the scores assigned to each human artwork and its small and large model versions.
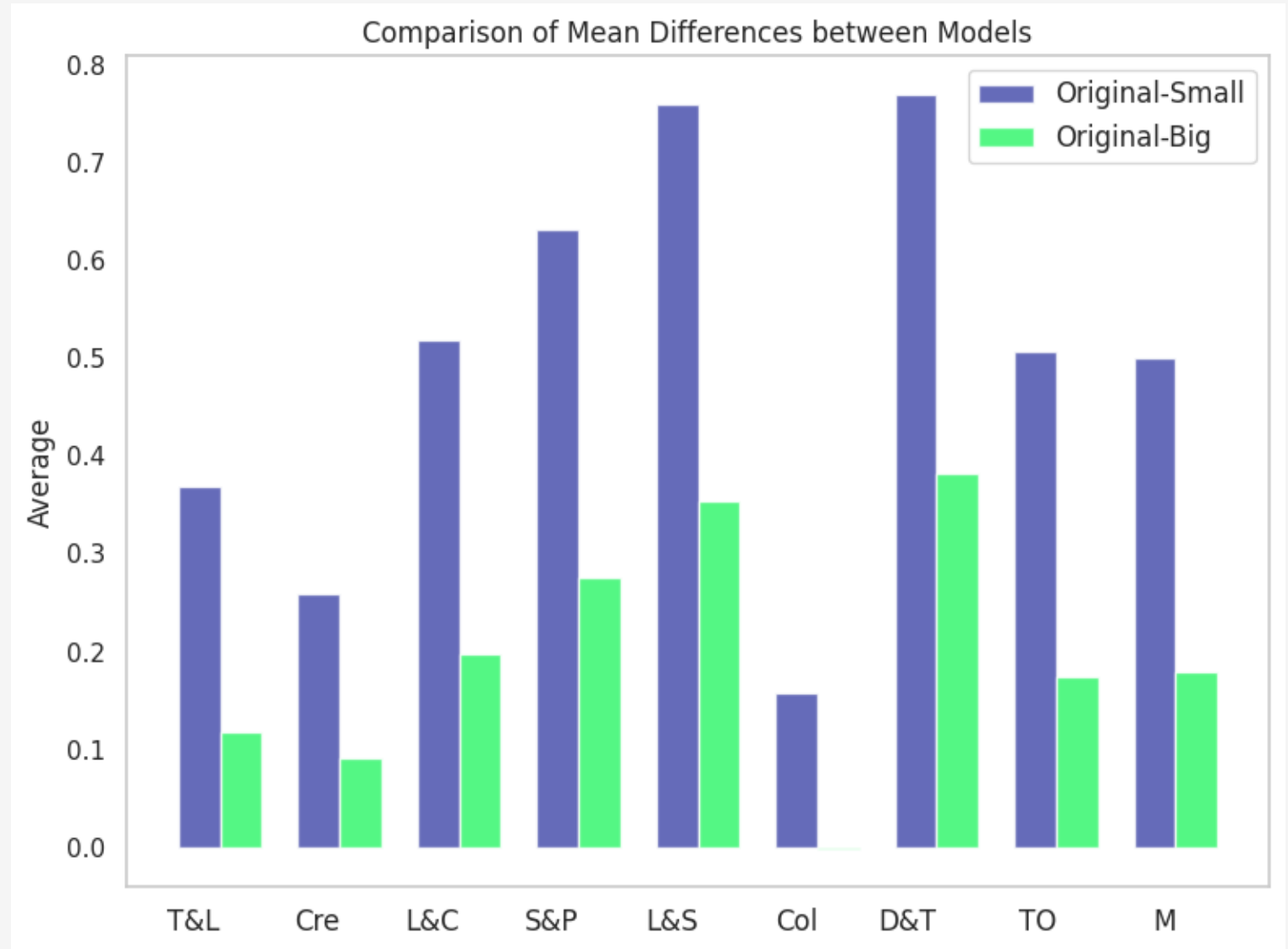
Results
Consistency

- Big model -> Closer to Human

- Scaling up model size **benefits all aspects** of aesthetic alignment **equally.**

Comparison of Mean Differences between Models
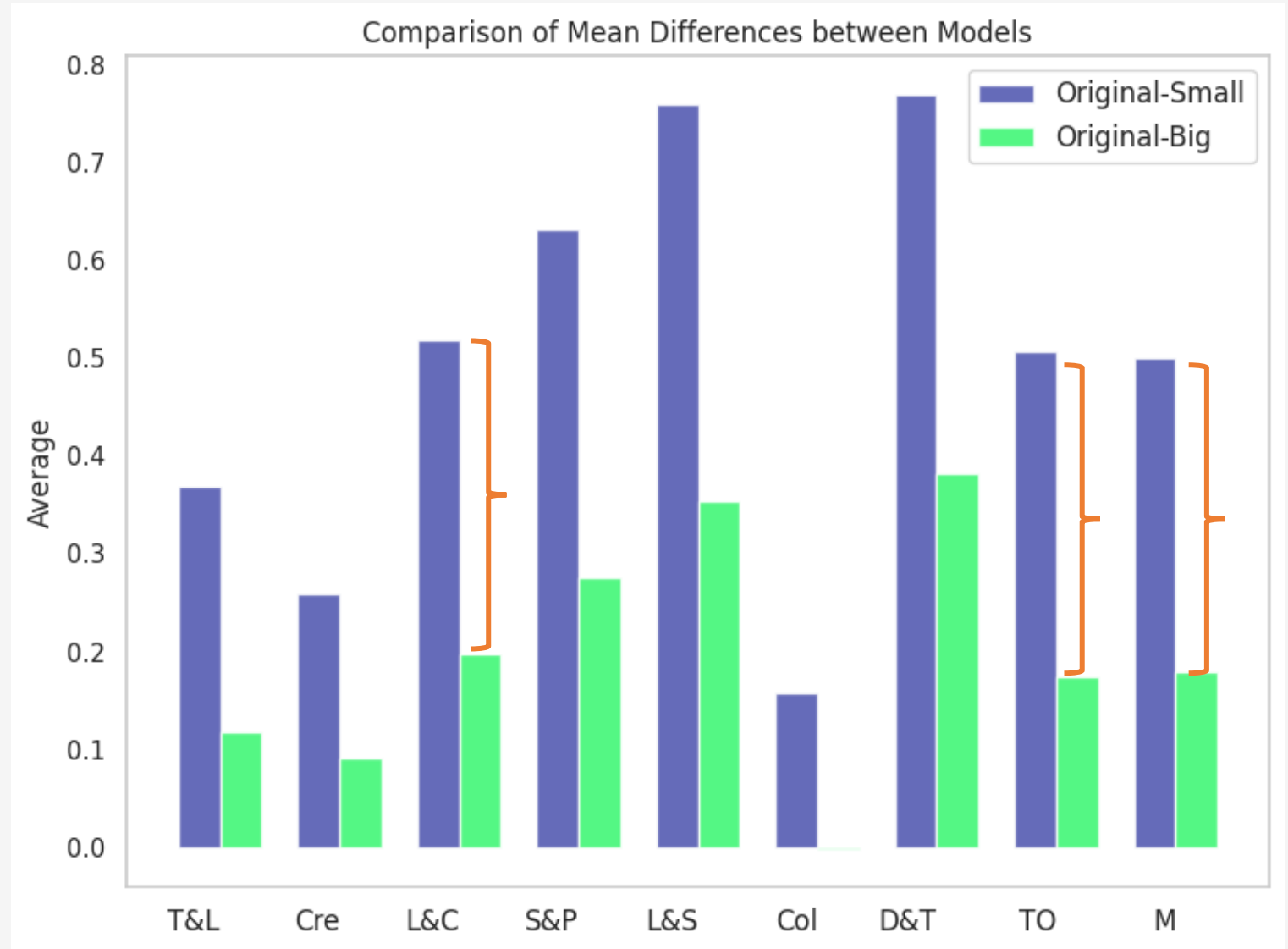
Results
Consistency

- Janus-Pro-7B achieves a **near-perfect** match in the **Color attribute**.

- Suggests strong ability to **replicate human-like color** usage when using auto-generated image descriptions from APDDv2.

Comparison of Mean Differences between Models
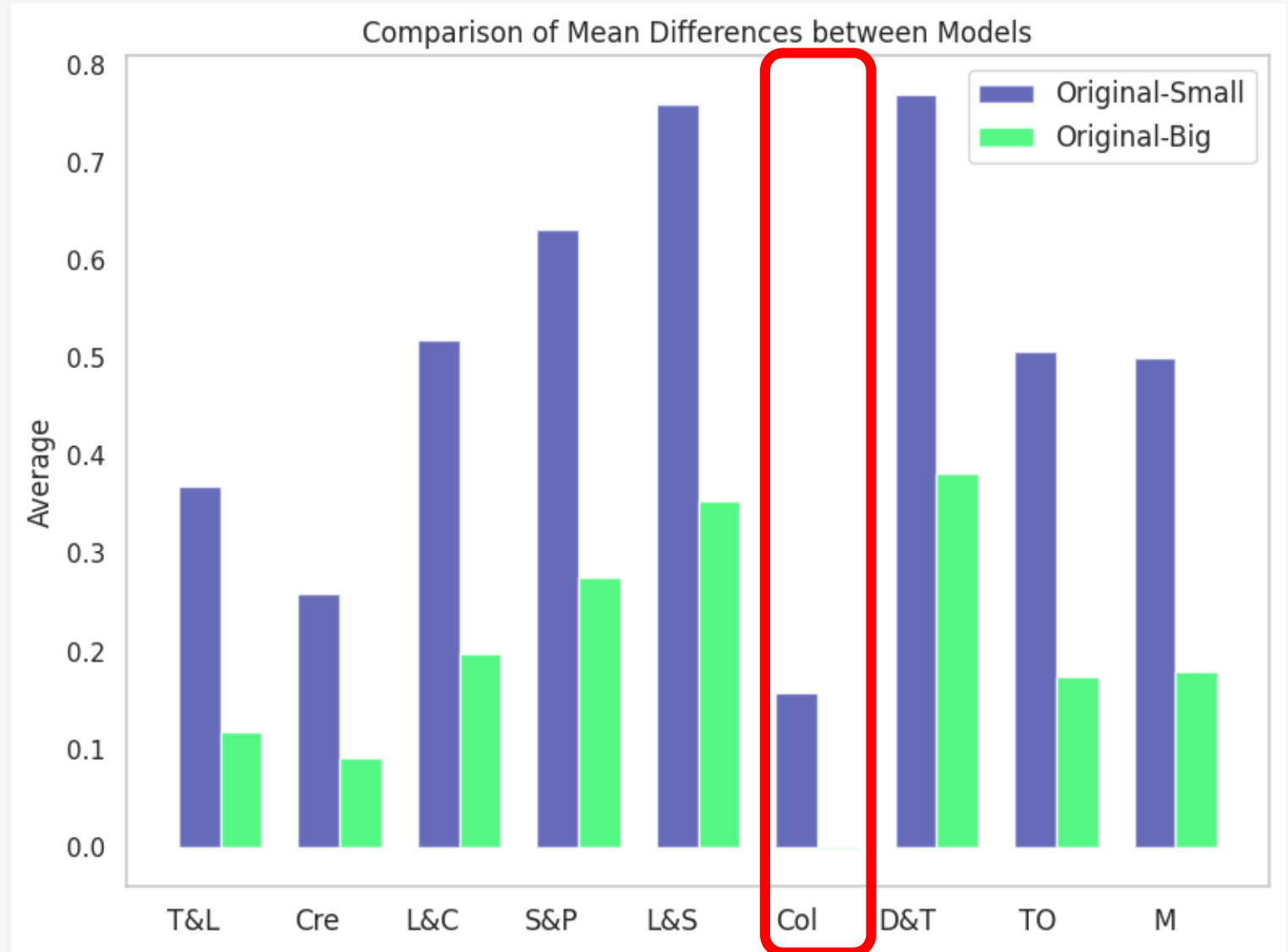
Takeaway

- Our main **contributions** are:
  - A **paired dataset:** human artworks and AI-generated images sharing the same textual description.
  - ArtCLIP **successfully distinguishes** human-created from AI-generated artworks in an **objective** and **consistent** way.
  - Comparison between generative **models of different sizes**.
  - **Larger models** produce images with aesthetic qualities **more aligned with human** artworks.
- ArtCLIP can serve as an automated and scalable **metric** for evaluating computational creativity.
- **Next steps:** qualitative evaluation with **human users**, and application to other **artistic domains**.

Questions?

Larissa D. Gomide

larissa.gomide@dcc.ufmg.br

Project IAIA-Inct on AI
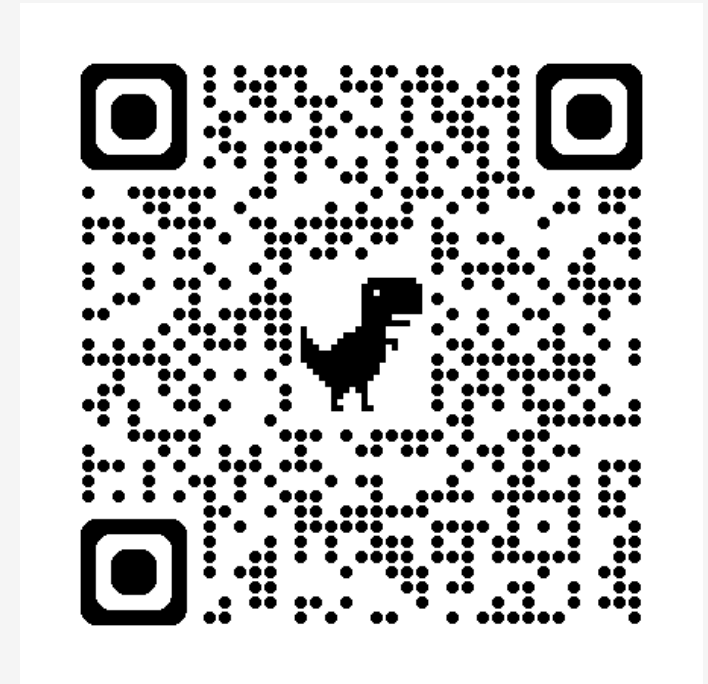
Let's keep the conversation going on LinkedIn.