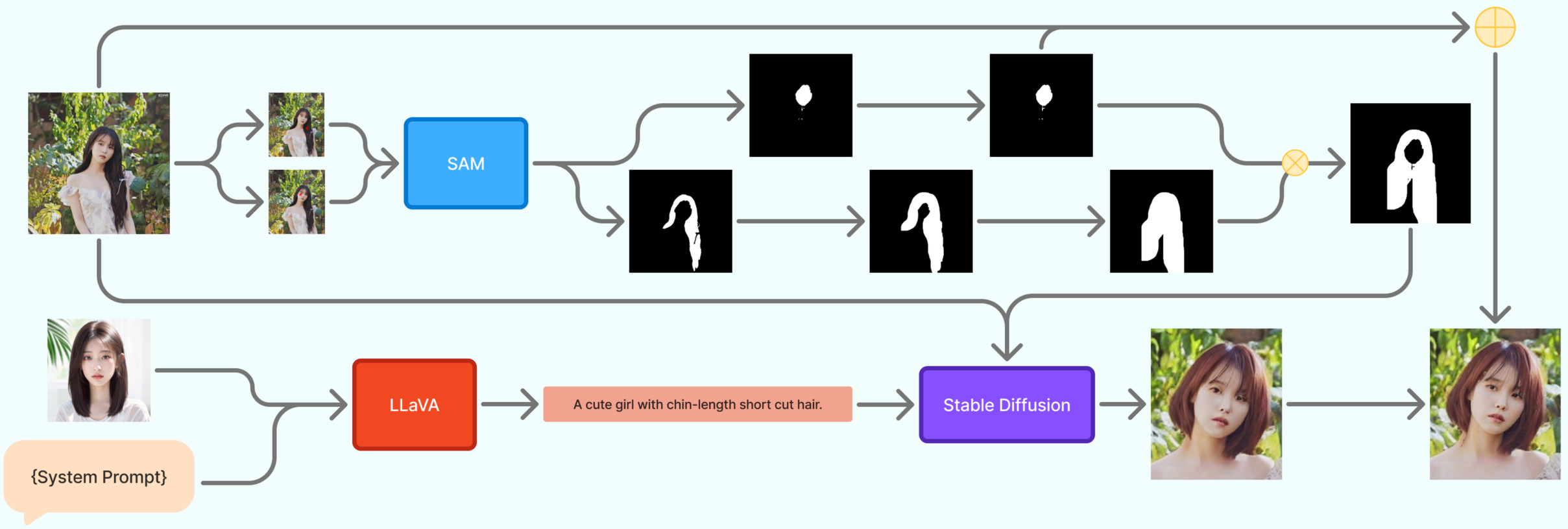
INTERACTIVE FACE STYLE EDITING WITH SAM2 AND STABLE DIFFUSION

CHENG-LIANG CHI, ZI-HUI LI, TING-WAN CHANG



Introduction

Hair editing in images is a challenging task due to the need for precise localization and natural-looking generation. Traditional methods require manual segmentation or coarse editing tools, while text-guided generative models often lack spatial control.

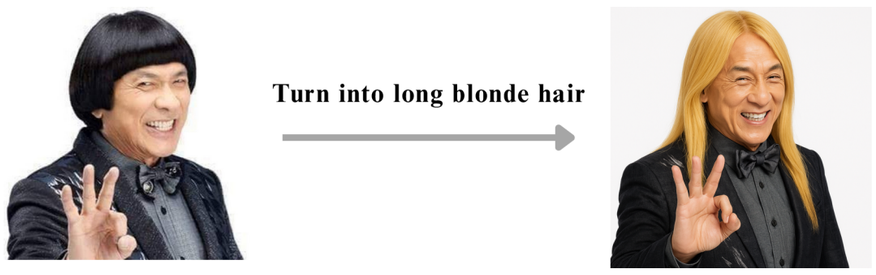
This project introduces an interactive pipeline focused exclusively on hairstyle editing. It combines:

- SAM2 for accurate segmentation of facial and hair regions,
- LLaVA for fine-grained text understanding,
- Stable Diffusion for high-quality visual generation.

Together, these components form a system that supports intuitive hairstyle modification using natural language prompts.

Motivation

While generative models like ChatGPT and Stable Diffusion excel at image creation, they lack spatial control, making localized editing such as targeting hair difficult. To solve this, we propose a hairstyle editing framework that uses segmentation masks and guided inpainting to constrain the generation region. By combining Stable Diffusion with spatial priors and natural language input, we enable intuitive, high-quality hair editing.



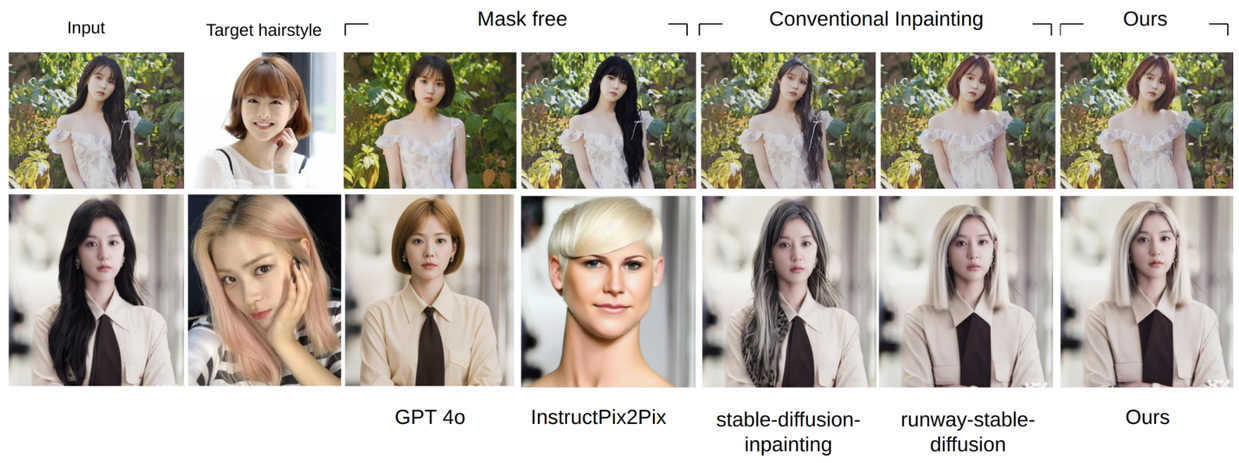
Result from GPT 4o

However, users often struggle to describe hairstyles in a way that models can understand. To simplify this, we use a Vision-Language Model (VLM) to automatically generate clear, detailed prompts from the input image.

Method

1. Segmentation (SAM2)
   - SAM2 segments face and hair regions.
2. Hair Region Expansion
   - Hair mask is expanded using directional dilation (e.g., vertical for long hair).
   - Face mask is subtracted to isolate the hairstyle area.
3. Image Understanding (LLaVA)
   - The description is used as a prompt for image generation.
4. Image Generation (Stable Diffusion)
   - The prompt, hairstyle mask, and original image are passed to Stable Diffusion and only the hairstyle region is modified.
5. Post-Processing
   - The original face is restored to maintain identity and consistency.

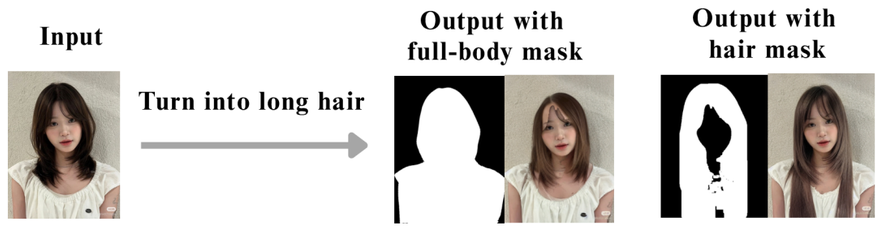
Results



Evaluation

	InstructPix2Pix	runway-stable-diffusion	stable-diffusion-inpainting	Ours	Original
CLIP	**0.2848**	0.2601	0.2808	0.2644	0.2334
FID	414.09	295.35	430.25	**293.20**	0
PSNR	15.22	15.35	12.85	**15.38**	inf
SSIM	0.6124	0.6776	0.5288	**0.6822**	1.0

Ablation study to different mask



Discussion

Compared to existing methods like InstructPix2Pix, our approach resolves two critical issues:

- Spatial Precision: Edits are limited strictly to hair regions using SAM2 and mask subtraction, avoiding global image changes.
- Text Robustness: LLaVA enables understanding of nuanced or vague hairstyle descriptions, even across different languages.

This design provides:

- More controllable outputs,
- Higher semantic alignment with the prompt,
- Better realism and identity preservation.

Although Stable Diffusion produces visually appealing results, it struggles to fully interpret complex hairstyle prompts. To improve control and user satisfaction, we propose two future directions:

- **Sketch-guided generation** – allowing users to draw simple hairstyle outlines or extracting contour sketches from images to provide clearer visual guidance for the model;
- **Iterative refinement with feedback** – incorporating interactive prompts (e.g., "longer", "curlier", "redo") to progressively refine outputs through user feedback.