Supplementary Material for ASTRA

Anonymous Submission

1 RAG-Pose Database and Training Data Pipeline

In this section, we provide a comprehensive description of our data generation and curation pipeline, which is crucial for both the Retrieval-Augmented Pose (RAG-Pose) system and the training of the ASTRA model. We first detail the systematic process of building the high-quality text-pose database . Then, we explain the pipeline for constructing the final training triplets.

1.1 High-Fidelity Text-Pose Database Construction

To overcome the data scarcity of complex poses, we built a large-scale, curated database linking descriptive text prompts to high-fidelity 2D pose skeletons. This process involves three main stages: systematic prompt engineering, VLM-driven image generation and curation, and finally, pose extraction and indexing.

Role:
You are a creative assistant for generating diverse scenarios for a text-to-image model. Your goal is to create rich and varied descriptive prompts for a given human action.

Rules:
1. You will be given a base 'action', like "a person stretching" or "two people dancing".
2. Generate 30 detailed scene descriptions for this action. Each description must be unique.
3. Vary the descriptions across these axes:
- Subject Attributes: age, body type, clothing style (e.g., "an elderly man", "a graceful ballerina").
- Action Specifics: mood or manner of the action (e.g., "stretching lazily", "dancing energetically").
- Environmental Context: location and setting (e.g., "on a misty mountain top", "in a crowded subway car").
- Camera Perspective: camera angle or shot type (e.g., "low-angle shot", "seen from above").
4. Each description should be a single, complete sentence. Do not exceed 77 tokens.

Example:
`[action]:` a person reading a book
`[SceneDescription1]:` A young woman is peacefully reading a thick novel in a sun-drenched, cozy armchair by a window.
`[SceneDescription2]:` An old man with glasses perches on a park bench, completely engrossed in a newspaper, seen from a distance.
`[SceneDescription3]:` A student frantically skims a textbook under the dim light of a desk lamp in a messy dorm room, high-angle shot.
... (Up to [SceneDescription30])

Figure 1: The system prompt used to instruct an LLM to generate diverse and detailed scene descriptions for our text-pose database, forming the foundation of our prompt engineering process.

Systematic and Diverse Prompt Engineering

To ensure the diversity and richness of our pose database, we began with a foundational set of over 300 human actions. We then employed a Large Language Model (LLM) to expand this set into over 9,000 unique prompts. This was achieved by instructing the LLM to generate variations along specific descriptive axes: subject attributes, action specifics, environmental context, and camera perspective. Figure 1 shows the system prompt used to guide the LLM for generating diverse scene descriptions for a given action.

VLM-Driven Generation, Curation, and Filtering

Using the 9,000+ generated prompts, we utilized FLUX.1-pro to synthesize an initial image pool. To ensure high quality, we developed a VLM-driven curation pipeline powered by GPT-4o. This pipeline filters images based on their semantic alignment with the prompt. We employed a multi-step Chain-of-Thought (CoT) prompting strategy to elicit a detailed and reliable evaluation from the VLM, as shown in Figure 2. This structured approach forces the VLM to first analyze the image components before making a final judgment, improving accuracy. The final alignment score S is computed as a weighted sum of three dimensions: subject consistency (s_1), interaction logic (s_2), and detail fidelity (s_3), with weights $\omega_1 = 0.45, \omega_2 = 0.35, \omega_3 = 0.20$. The weights ($\omega_1 = 0.45, \omega_2 = 0.35, \omega_3 = 0.20$) were optimized via regression on a held-out set against human preferences to prioritize subject and interaction correctness. Only images with $S > 0.85$ were accepted.

Role:
You are a meticulous AI assistant for evaluating the quality and alignment of an image generated from a text prompt.

Input Format:
You will receive a [Prompt] and an [Image].

Step 1: Analysis
First, describe the key elements in the [Image]. Identify the subjects, their actions, and the environment. Do not make any judgments in this step.

Step 2: Component-wise Comparison
Now, compare the image to the [Prompt] based on the following criteria. Describe any discrepancies in detail.
- Subject Consistency: Does the number, appearance, and type of subjects in the image match the prompt?
- Interaction Logic: Is the action or interaction depicted physically plausible and consistent with the prompt?
- Detail Fidelity: Are other details like the environment, style, and composition accurately rendered?

Step 3: Scoring
Based on your detailed comparison in Step 2, provide a numerical score from 0.0 (no alignment) to 1.0 (perfect alignment) for each of the three criteria. Output the final result in a JSON format ONLY.
Output Format:

```
{
"justification_s1": "...", "score_s1": 0.0-1.0,
"justification_s2": "...", "score_s2": 0.0-1.0,
"justification_s3": "...", "score_s3": 0.0-1.0
}
```

Figure 2: The Chain-of-Thought (CoT) prompt used to guide GPT-4o for the semantic verification and filtering of generated images. This multi-step process ensures a thorough and reliable quality assessment.

1.2 Training Data Construction Pipeline

For training ASTRA, we require data triplets of $\{prompt, ref\_imgs, pose\}$. We designed an automated pipeline to construct these triplets from a set of high-quality target images (which can be from our curated database or other sources like COCO). This pipeline, illustrated in Figure 3, ensures that the training data has strong internal consistency between identity, structure, and text. Unlike some methods, which uses cropped images and faces copy-paste issues, our use of Flux.1 Kontext to re-edit and generate new reference images from segmented subjects effectively mitigates this problem by creating novel, clean reference views of the subject.

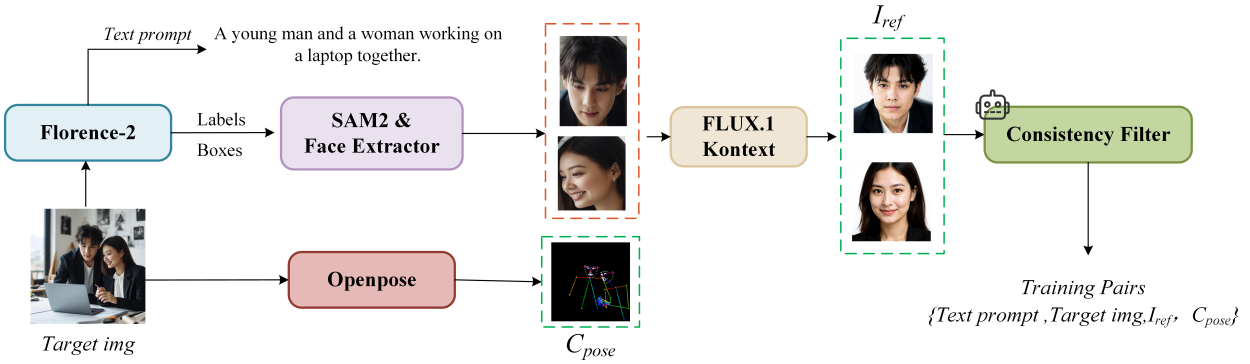

Figure 3: Our automated pipeline for constructing training triplets. (1) A high-quality target image is selected. (2) Florence-2 generates a descriptive caption (the 'prompt') and detects subject bounding boxes. (3) SAM2 produces precise segmentation masks for each subject. (4) The segmented subjects are fed into Flux.1 Kontext, which generates diverse, clean reference images. (5) OpenPose extracts the pose skeleton from the original target image. This process avoids the copy-paste problem and yields high-quality, consistent training data.

2 Analysis of Key Hyperparameters

In this section, we provide an analysis of several key hyperparameters that influence the performance and efficiency of ASTRA. We focus on the LoRA rank, which directly impacts the model's capacity, and other crucial parameters from our RAG-Pose pipeline.

2.1 Analysis of LoRA Rank

We conducted an ablation study on the rank of the LoRA adapter used for training ASTRA. The rank determines the number of trainable parameters and the expressive power of the adapter. We scaled the rank from 32 to 512 and evaluated performance on our COCO-based complex pose benchmark. As shown in Figure 4, increasing the rank from 32 to 256 yields significant improvements in both identity preservation (CLIP-I) and pose accuracy (OKS). Beyond a rank of 256, the performance gains begin to saturate, with only marginal improvements observed at rank 512. The increase in training memory and time becomes more pronounced at higher ranks. Considering the trade-off between performance and computational cost, we selected a rank of **512** for our final model to maximize performance, but a rank of 256 offers a compelling balance for more resource-constrained scenarios.

2.2 Discussion of Other Hyperparameters

Besides the LoRA rank, several other hyperparameters are critical to the success of our framework.

- **Retrieval Confidence Threshold (α_u):** This threshold in the RAG-Pose pipeline determines whether a retrieved pose is used. We set $\alpha_u = 0.55$ based on empirical evaluation on a validation set.As shown in Figure 5, a lower value (e.g., 0.4) increased the retrieval rate but often introduced irrelevant or poorly matched poses, degrading the final OKS score. A higher value (e.g., 0.7) led to overly conservative retrieval, causing the model to default

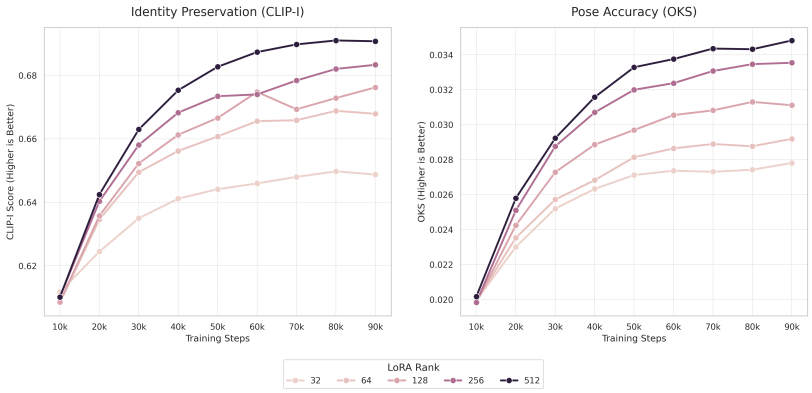

Figure 4: Performance of ASTRA on the COCO-based benchmark with varying LoRA ranks. Both CLIP-I and OKS scores improve steadily with increasing rank, with performance plateauing after rank 256. We chose rank 512 for the best possible results.

to text-only guidance too frequently, which also hurt performance on complex prompts. The chosen value of 0.55 provided the best balance between precision and recall for pose retrieval.

- **Learning Rate:** We found that a learning rate of 1×10^{-5} was optimal. A higher rate (e.g., 2×10^{-5}) led to unstable training and divergence, particularly due to the direct injection of clean conditioning tokens into the pre-trained model. A lower rate (e.g., 5×10^{-6}) resulted in excessively slow convergence without significant performance benefits.

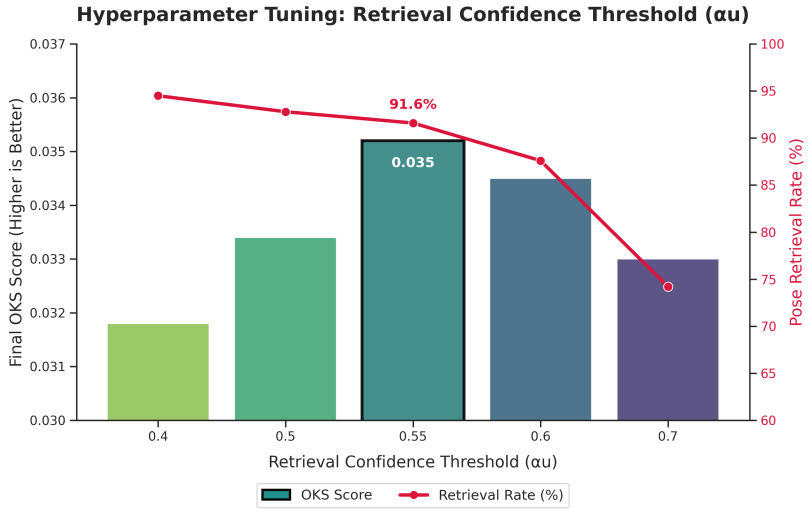

Figure 5: Performance of ASTRA on the COCO-based benchmark with varying α_u.

3 Construction of the COCO-based Complex Pose Benchmark

To evaluate ASTRA's ability to generate complex, multi-person scenes with precise poses, we constructed a specialized benchmark based on the COCO 2017 Keypoints dataset. The construction and evaluation process is designed to be both comprehensive and fair to all models.

System Role:
You are a highly precise human pose analyst. Your task is to generate a single, detailed descriptive sentence for an image, focusing exclusively on the poses and actions of the person(s).
Instructions:
1. Analyze the input [Image].
2. Describe the primary action (e.g., "running," "sitting," "reaching").
3. Detail the orientation of the torso and head (e.g., "leaning forward," "facing left").
4. Specify the position of arms and legs with as much detail as possible (e.g., "with the right arm extended upwards and the left leg bent at a 90-degree angle").
5. If there are multiple people, describe their relative positions and interactions.
6. Combine all details into one comprehensive sentence.
Example Output:
`[Description]:` A person is lunging forward to their right, with their right arm fully extended to catch a frisbee and their left leg bent behind them for balance.

Figure 6: The system prompt used to instruct a VLM to generate pose-centric descriptions from images. This forms the textual pose prior for all models in our benchmark.

Benchmark Construction Pipeline

The process begins by selecting challenging images from the dataset containing two or more persons. A crucial step is the creation of a pose-centric text prompt for each image. To achieve this, we use a Vision-Language Model (VLM) guided by the specific instructions shown in Figure 6, which forces the VLM to generate a highly detailed description of body orientation, limb positions, and actions. This detailed prompt serves as the primary textual guidance for the target pose. To ensure a fair comparison, all models under evaluation receive this same detailed prompt. This methodology provides every model with a strong textual pose prior, thereby mitigating potential biases that could arise from ASTRA's unique RAG-based retrieval mechanism and ensuring the evaluation truly reflects each model's ability to interpret complex pose descriptions. Concurrently, we use Florence-2's object detection to identify subject bounding boxes, which are then passed to SAM2 to segment and extract clean, background-free reference images (`ref_imgs`).

Evaluation Protocol

During evaluation, the original COCO image serves as the ground truth. A model is tasked with generating the scene given the pose-centric `prompt` and the `ref_imgs`. We assess the output using a suite of metrics. The primary metric for pose accuracy is the **Object Keypoint Similarity (OKS)**, which is calculated as:

$$OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2) \cdot \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{1}$$

where d_i is the Euclidean distance between the detected and ground truth keypoint, s is the object's scale, k_i is a per-keypoint constant controlling falloff, and $\delta(v_i > 0)$ indicates if the keypoint is labeled in the ground truth. In addition to OKS, we evaluate identity preservation with CLIP-I, semantic alignment between the image and prompt with CLIP-T, and low-level feature similarity with DINO. This robust framework allows for a thorough and fair measurement of joint identity and pose control.

4 User Study Evaluation

To complement our quantitative metrics and assess the quality of our results, we conducted a comprehensive user study. We recruited 36 evaluators with varying levels of expertise and presented them with 400 test cases, each containing

a text prompt and corresponding reference images. In a randomized comparison, they evaluated images generated by ASTRA against other leading models. Evaluators were asked to select the best result based on five criteria: Subject Similarity, Action Fidelity , Text Fidelity(Background), Text Fidelity(Subject), Compositional Quality, and Overall Visual Appeal.

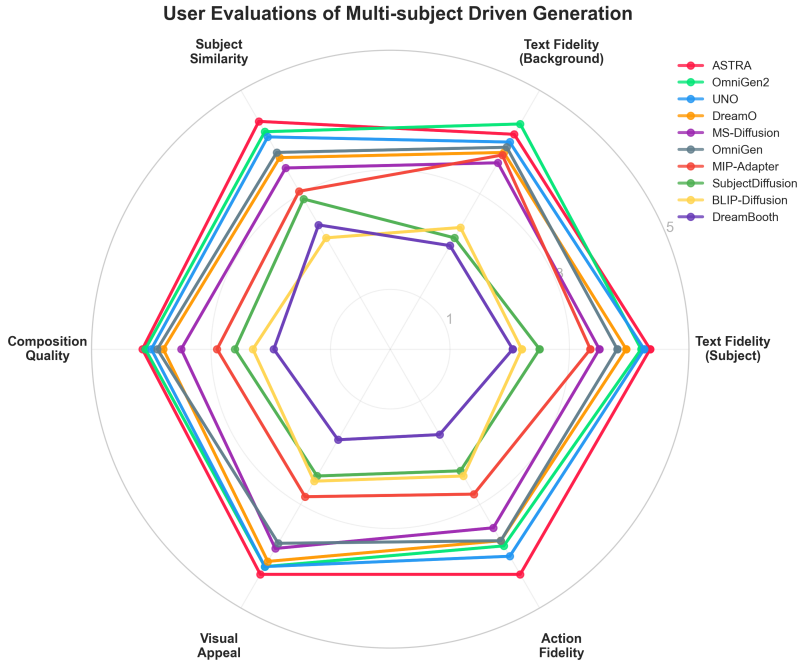

Figure 7: Radar chart of user evaluation of methods for multi-subject driven generation on different dimensions

As summarized in Figure 7, the results show a clear and consistent preference for ASTRA. Our method significantly outperformed all baselines, receiving the highest scores across most dimensions. Notably, its superiority was most pronounced in Action Fidelity and Subject Similarity, which directly validates the effectiveness of our RAG-Pose system and Disentangled Semantic Modulation (DSM) module. This user study confirms that the quantitative improvements we measured are perceptually significant, as human evaluators found ASTRA's outputs to be more accurate in both pose and identity, leading to a higher overall visual quality.

5 Application Scenarios

ASTRA's unique ability to control pose and identity with high fidelity opens up a range of practical and creative applications. Its strong generalization,as shown in Figure 8, built upon the FLUX.1 foundation and our ASTRA framework, allows it to excel even in scenarios not explicitly seen during training.

Pre-visualization for Creative Industries. For filmmakers, animators, and choreographers, ASTRA serves as a powerful pre-visualization tool. A director can generate a specific shot by providing a prompt like *"a detective cautiously peeking around a corner, low-angle shot"* and a reference image of the actor. The RAG-Pose system can retrieve a precise skeleton for "peeking around a corner," allowing for rapid iteration on storyboards and animatics with the correct actor and composition.

Pose-Specific Virtual Try-on. ASTRA advances virtual try-on by enabling dynamic poses. Instead of static, A-pose models, users can see how clothing looks in realistic scenarios described by prompts like *"a woman jogging in the park wearing a new jacket"* or *"a man doing a yoga warrior pose"*. This provides a much more informative and engaging user experience.

Enhanced Storytelling and Character Generation. Content creators for games, comics, or marketing can use ASTRA to generate consistent characters in a multitude of narrative situations. By preserving identity and precisely

controlling body language, ASTRA can depict a character's emotional journey—from *"a hero standing triumphantly on a cliff"* to *"the same hero sitting defeated in the rain"*.

Stylized Generation with Controlled Poses. While inheriting the powerful stylization capabilities of FLUX.1, ASTRA ensures that artistic expression does not come at the cost of structural control. It can generate images in various styles (e.g., *"oil painting of two people waltzing," "anime-style drawing of a character casting a spell"*) while faithfully maintaining the specified pose and identity, a task that is challenging for most models.

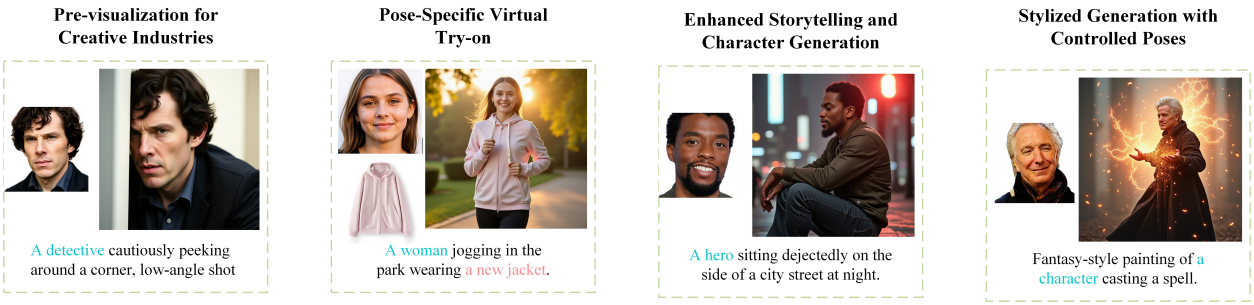

Figure 8: Diverse application scenarios enabled by ASTRA.

6 Limitation and Future Work

Although ASTRA demonstrates state-of-the-art capabilities in multi-subject generation with precise pose control, our approach is primarily centered on retrieval-augmented structural guidance. The performance of our model is therefore linked to the scope and quality of the pre-curated pose database, which currently limits its ability to generate highly novel or out-of-distribution poses not represented within it. To address this core limitation, we plan to evolve our framework towards a more flexible hybrid "retrieve-or-generate" paradigm. This would involve **incorporating a lightweight generative module** capable of creating novel pose structures directly from textual descriptions or simple user sketches, **providing a robust fallback when a suitable pose cannot be retrieved**.

Beyond overcoming this constraint, this hybrid philosophy opens the door to expanding ASTRA's capabilities over just poses. We envision **extending our database and retrieval mechanism to include other conditional modalities**, such as compositional layouts and artistic exemplars. The core architectural strengths of ASTRA are uniquely suited for this advancement; the asymmetric positional encoding of EURoPE could enforce spatial arrangements from a retrieved layout, while the Disentangled Semantic Modulation (DSM) could be adapted to inject stylistic information as another controllable dimension, separate from identity. By integrating both retrieval and generation for these various elements, and **diversifying our training data** accordingly, we can transform ASTRA into a more holistic generative framework that bridges the gap between the reliability of retrieval and the creative freedom of generation.