A. Prompts

We show the system and user prompts used for VLM judges in different scenarios. We then show the way we organize specialists' output in the judge prompt. Last but not least, the subsequent subsection shows prompts used for sample-wise automatic tool selection.

A.1. VLM-as-a-judge Prompts

Image Editing (pairwise)
System prompt:

```
You are an image evaluation expert. You
    will be presented with an editing
    request, a source image, and two edited
    images.

Your task is:
   (1) Analyze the edited images based on
       multiple aspects, such as request
       fulfillment (i.e. text-image
       alignment), technical quality,
       aesthetic appeal, detail preservation
       , etc.
   (2) On each aspect or the final judgment
       , conclude with either
      * "image1" if the first edited image
          is better in the aspect or overall
      * "image2" if the second edited image
           is better in the aspect or
          overall

It's recommended to focus on the most
    important aspects, e.g., request
    fulfillment, technical quality, detail
    preservation, etc., to make the decision
     which image is better overall. They
    will vary case by case.


### Additional Notes:

To support your evaluation, you may be
    provided with assessments of each edited
     image by other specialists, which you
    should consider in your analysis. Based
    on the request and the output images,
    please attend to the most relevant
    judgments of the specialists, while you
    can ignore other judgments. You should
    have your own stance on the assessment
    of the images.


### Response Format:

Format your response into XML tags as shown
     below:

<evaluation>
   <image1>Your concise thoughts about
       edited version 1</image1>
   <image2>Your concise thoughts about
       edited version 2</image2>
   <alignment>"image1" or "image2" which is
        better in term of request
       fulfillment</alignment>
   <quality>"image1" or "image2" which is
       better in term of image quality (
       technical quality, aesthetic appeal,
       etc.) </quality>
   <preservation>"image1" or "image2" which
        is better in term of detail
       preservation</preservation>
   <analysis>Your short overall analysis to
        justify your judgment</analysis>
   <judgment>Conclude with either "image1"
       if the first edited image is better
       or "image2" if the second edited
       image is better</judgment>
</evaluation>
```

User prompt:

```
[
   {"type": "text",
    "text": f"Given the following image
        editing prompt, original image, and
        two edited images:\n\nUser
        instruction: {edit_request}"},
   {"type": "text", "text": f"\n\nOriginal
       image: "},
   {"type": "image_url", "image_url": {"url
       ": original_image}},
   {"type": "text", "text": f"\n\nEdited
       image 1: "},
   {"type": "image_url", "image_url": {"url
       ": edited_image1}},
   {"type": "text", "text":
       specialist_assessments_1},
   {"type": "text", "text": f"\n\nEdited
       image 2: "},
   {"type": "image_url", "image_url": {"url
       ": edited_image2}},
   {"type": "text", "text":
       specialist_assessments_2},
   {"type": "text", "text": "\n\nNow,
       please compare the edited images in
       term of image quality, their
       alignment with the request, and
       detail preservation."}
]
```

Image Generation (pairwise)
System prompt:

```
You are an image evaluation expert. You
    will be presented with a generation
    request and two generated images.

Your task is:
   (1) Analyze the generated images based
       on multiple aspects, such as request
       fulfillment (i.e. text-image
       alignment), technical quality,
       aesthetic appeal, etc.
   (2) On each aspect or the final judgment
       , conclude with either
      * "image1" if the first generated
          image is better
      * "image2" if the second generated
          image is better

It's recommended to focus on the most
    important aspects, e.g., request
    fulfillment, technical quality, etc., to
     make the decision which image is better
     in overall. They will vary case by case
    .


### Additional Notes:

To support your evaluation, you may be
    provided with assessments of each
    generated image by other specialists,
    which you should consider in your
    analysis. Based on the request and the
    output images, please attend to the most
     relevant judgments of the specialists,
    while you can ignore other judgments.
    You should have your own stance on the
    assessment of the images.


### Response Format:

Format your response into XML tags as shown
     below:
<evaluation>
   <image1>Your concise thoughts about
       generated version 1</image1>
   <image2>Your concise thoughts about
       generated version 2</image2>
   <alignment>"image1" or "image2" which is
        better in term of request
       fulfillment</alignment>
   <quality>"image1" or "image2" which is
       better in term of image quality (
       technical quality, aesthetic appeal,
       etc.) </quality>
   <analysis>Your short overall analysis to
        justify your judgment</analysis>
   <judgment>Conclude with either "image1"
       if the first generated image is
       better or "image2" if the second
       generated image is better</judgment>
</evaluation>
```

User prompt:

```
[
   {"type": "text",
    "text": f"Given the following generation
         prompt and two generated images:\n\
        nUser instruction: {
        generation_request}"},
   {"type": "text", "text": f"\n\nGenerated
        image 1: "},
   {"type": "image_url", "image_url": {"url
       ": image1}},
   {"type": "text", "text":
       specialist_assessments_1},
   {"type": "text", "text": f"\n\nGenerated
        image 2: "},
   {"type": "image_url", "image_url": {"url
       ": image2}},
   {"type": "text", "text":
       specialist_assessments_2},
   {"type": "text", "text": "\n\nNow,
       please compare the generated images
       in term of image quality and their
       alignment with the request."}
]
```

Image Generation (pointwise)

System prompt:

```
You are an image evaluation expert. You
    will be presented with a generation
    request and a generated image.

Your task is:
   (1) Analyze the generated image based on
        multiple aspects, such as request
       fulfillment (i.e. text-image
       alignment), technical quality,
       aesthetic appeal, etc.
   (2) On each aspect or the final judgment
       , conclude with a score in a Likert
       -10 scale, with 1 being the lowest
       point, and 10 being the highest point
       . Indeed, the final judgment should
       reflect your evaluation of the
       generated image in each aspect.


### Additional Notes:

To support your evaluation, you may be
    provided with assessments of the
    generated image by other specialists,
```

```
    which you should consider in your
    analysis. Based on the request and the
    output images, please attend to the most
     relevant judgments of the specialists,
    while you can ignore other judgments.
    You should have your own stance on the
    assessment of the image.


### Response Format:

Format your response into XML tags as shown
     below:
<evaluation>
   <image>Your thoughts about the generated
        image</image>
   <alignment>A score in the range of 1 to
       10 on request fulfillment</alignment>
   <quality>A score in the range of 1 to 10
        on image quality</quality>
   <judgment>A overall score in the range
       of 1 to 10</judgment>
</evaluation>
```

User prompt:

```
[
   {"type": "text",
    "text": f"Given the following generation
         prompt and the generated image:\n\
        nUser instruction: {
        generation_request}"},
   {"type": "text", "text": f"\n\nGenerated
        image: "},
   {"type": "image_url", "image_url": {"url
       ": image}},
   {"type": "text", "text":
       specialist_assessments},
   {"type": "text", "text": "\n\nNow,
       please analyze the generated image in
        term of image quality and its
       alignment with the request."}
]
```

Specialist Assessments: We organize specialists' output in bullet points and in input order. Also, before each specialist's output, we have a short description of the tool. For example, the output of AesExpert, PAL4VST, and Q-SiT will be organized as follows:

```
**1. Aesthetic description**: <description>
**2. Perceptual artifact ratio from 0-1,
    where 0 means no artifacts**: <ratio>
**3. Overall quality score from 0-1, where
    1 is perfect**: <score>
```

A.2. Automatic Tool Selection Prompts

Image Editing (pairwise)

System prompt:

```
You are an image evaluation expert. You
    will be presented with an editing
    request, a source image, and two edited
    images.

Later, you will 1) analyze the edited
    images based on multiple aspects, such
    as request fulfillment (i.e., text-image
     alignment), technical quality,
    aesthetic appeal, detail preservation,
    etc., then 2) compare the two edited
    images and determine which one is better
     overall.

Multiple specialist models can provide
    assessments of the edited images in low-
    level details. One specialist may or may
     not be helpful in this case, as it may
    not be a significant factor, or you may
    already possess that capability. The
    following are the specialists available
    for you to choose from:

* AesExpert: Analyzes aesthetic qualities
    of images and provides a detailed
    aesthetic description.
* DeepFace: Measures the similarity
    distance between faces in the source
    image and the edited image.
* DreamSim: Measures the structural
    similarity distance between the source
    image and the edited image.
* GroundingDINO: Detects and counts objects
     in images to verify prompt following.
* PAL4VST: Computes perceptual artifact
    ratio in images. Value from 0-1, where 0
     means no artifacts.
* QSiT: Calculates an overall quality score
     for the images. Value from 0-1, where 1
     means best quality.


### Main task

Your task now is to select a minimal set of
     tools you need for the case, based on
    your observation of the input (both
    prompt and image) and your existing
    capability. Also, before concluding tool
     selection, please briefly discuss why
    you choose or do not choose each tool.

### Response Format:

Format your response into XML tags as shown
     below:
<evaluation>
```

```
   <discussion>Discuss your choices of
       tools</discussion>
   <tool>A minimal set of tools that you
       need for this evaluation case. Only
       return a string of tool names,
       separated by `_`, e.g.,
       AesExpert_QSiT</tool>
</evaluation>
```

User prompt:

```
[
   {"type": "text",
    "text": f"Given the following image
        editing prompt, original image, and
        two edited images:\n\nUser
        instruction: {edit_request}"},
   {"type": "text", "text": f"\n\nOriginal
       image: "},
   {"type": "image_url", "image_url": {"url
       ": original_image}},
   {"type": "text", "text": f"\n\nEdited
       image 1: "},
   {"type": "image_url", "image_url": {"url
       ": edited_image1}},
   {"type": "text", "text": f"\n\nEdited
       image 2: "},
   {"type": "image_url", "image_url": {"url
       ": edited_image2}},
   {"type": "text", "text": "\n\nNow,
       please briefly discuss why you choose
        or not choose each tool, then
       conclude tool selection."}
]
```

Image Generation (pairwise)

System prompt:

```
You are an image evaluation expert. You
    will be presented with a generation
    request and two generated images.

Later, you will 1) analyze the generated
    images based on multiple aspects, such
    as request fulfillment (i.e., text-image
     alignment), technical quality,
    aesthetic appeal, etc., then 2) compare
    the two generated images and determine
    which one is better overall.

Multiple specialist models can provide
    assessments of the generated images in
    low-level details. One specialist may or
     may not be helpful in this case, as it
    may not be a significant factor, or you
    may already possess that capability. The
     following are the specialists available
     for you to choose from:

* AesExpert: Analyzes aesthetic qualities
    of images and provides a detailed
    aesthetic description.
* GroundingDINO: Detects and counts objects
     in images to verify prompt following.
* PAL4VST: Computes perceptual artifact
    ratio in images. Value from 0-1, where 0
     means no artifacts.
* QSiT: Calculates an overall quality score
     for the images. Value from 0-1, where 1
     means best quality.


### Main task

Your task now is to select a minimal set of
     tools you need for the case, based on
    your observation of the input (both
    prompt and image) and your existing
    capability. Also, before concluding tool
     selection, please briefly discuss why
    you choose or do not choose each tool.

### Response Format:

Format your response into XML tags as shown
     below:
<evaluation>
   <discussion>Discuss your choices of
       tools</discussion>
   <tool>A minimal set of tools that you
       need for this evaluation case. Only
       return a string of tool names,
       separated by `_`, e.g.,
       AesExpert_QSiT</tool>
</evaluation>
```

User prompt:

```
[
   {"type": "text",
    "text": f"Given the following generation
         prompt and two generated images:\n\
        nUser instruction: {
        generation_request}"},
   {"type": "text", "text": f"\n\nGenerated
        image 1: "},
   {"type": "image_url", "image_url": {"url
       ": image1}},
   {"type": "text", "text": f"\n\nGenerated
        image 2: "},
   {"type": "image_url", "image_url": {"url
       ": image2}},
   {"type": "text", "text": "\n\nNow,
       please briefly discuss why you choose
        or not choose each tool, then
       conclude tool selection."}
]
```

Image Generation (pointwise)

System prompt:

```
You are an image evaluation expert. You
    will be presented with a generation
    request and a generated image.

Later, you will 1) analyze the generated
    image based on multiple aspects, such as
     request fulfillment (i.e., text-image
    alignment), technical quality, aesthetic
     appeal, etc., then 2) conclude with a
    preference score for the generated image
     in a Likert-10 scale, with 1 being the
    lowest point, and 10 being the highest
    point.

Multiple specialist models can provide
    assessments of the generated image in
    low-level details. One specialist may or
     may not be helpful in this case, as it
    may not be a significant factor, or you
    may already possess that capability. The
     following are the specialists available
     for you to choose from:

* AesExpert: Analyzes aesthetic qualities
    of images and provides a detailed
    aesthetic description.
* GroundingDINO: Detects and counts objects
     in images to verify prompt following.
* PAL4VST: Computes perceptual artifact
    ratio in images. Value from 0-1, where 0
     means no artifacts.
* QSiT: Calculates an overall quality score
     for the images. Value from 0-1, where 1
     means best quality.


### Main task

Your task now is to select a minimal set of
     tools you need for the case, based on
    your observation of the input (both
    prompt and image) and your existing
    capability. Also, before concluding tool
     selection, please briefly discuss why
    you choose or do not choose each tool.

### Response Format:

Format your response into XML tags as shown
     below:
<evaluation>
   <discussion>Discuss your choices of
       tools</discussion>
   <tool>A minimal set of tools that you
       need for this evaluation case. Only
       return a string of tool names,
       separated by `_`, e.g.,
       AesExpert_QSiT</tool>
</evaluation>
```

User prompt:

```
[
   {"type": "text",
    "text": f"Given the following generation
        prompt and the generated image:\n\
       nUser instruction: {
       generation_request}"},
   {"type": "text", "text": f"\n\nGenerated
       image: "},
   {"type": "image_url", "image_url": {"url
       ": image}},
   {"type": "text", "text": "\n\nNNow,
       please briefly discuss why you choose
        or not choose each tool, then
       conclude tool selection."}
]
```

B. Training Hyperparameters

Most of the hyperparameters were set using the default values provided in https://github.com/OpenGVLab/InternVL. We use training batch size of 256 because the training set is fairly small and finetuned for 3 epochs. We set the maximum number of dynamic patches to 6 to reduce memory demand. We train all components of the model including the vision encoder, projector, and the language model.

C. Further Analysis

We analyze judgment distributions of base VLMs and AUTORATER variants in Figure 5. For pairwise comparison, GPT-4o in PSR benchmark has significantly higher probability than random in predicting the winner as Image 1. For other cases (with ImageReward (pairwise) or with Gemini-2.5-Pro), the score distributions are relatively balanced. For pointwise rating, we can clear observe that GPT-4o and AUTORATER with GPT-4o tend to give high scores (8, 9, 10) more frequently than other scores on both ImageReward (pointwise) and AIGIQA-30K. Also, the distribution is not widely spreaded as Gemini-2.5-Pro and AUTORATER with Gemini-2.5-Pro. That partially explain the relatively higher Pearson correlations of GPT-4o in pointwise rating scenarios, as human ratings also tend to be high scores. Overall, AUTORATER variants have similar score distributions as their base VLMs. It indicates that the integration of specialists does not drastically alter the inherent scoring tendencies of the VLMs.

Finally, we analyze the tool selection results of AUTORATER with Auto tool selection mechanism in Figure 6. Notably, VLMs usually choose AesExpert and Q-SiT more frequently than other specialists across different evaluation scenarios. It suggests that aesthetic description from AesExpert and overall quality score from Q-SiT are generally useful for VLM judges to make decisions. GroundingDINOv2 is less frequently selected than AesExpert and Q-SiT, and interestingly, Gemini-2.5-Pro selectes GroundingDINOv2 2 to 3 times less than GPT-4o. It indicates that current advanced VLMs are sufficient for object detection and counting, thus there is no need for an external object detector. On the other hand, DeepFace is only applicable for image editing, and GPT-4o rarely use the tool. In fact, the majority of curated PSR benchmark do not involve in human identity. Thus, by skipping the tool, GPT-4o optimizes efficiency without sacrificing performance.

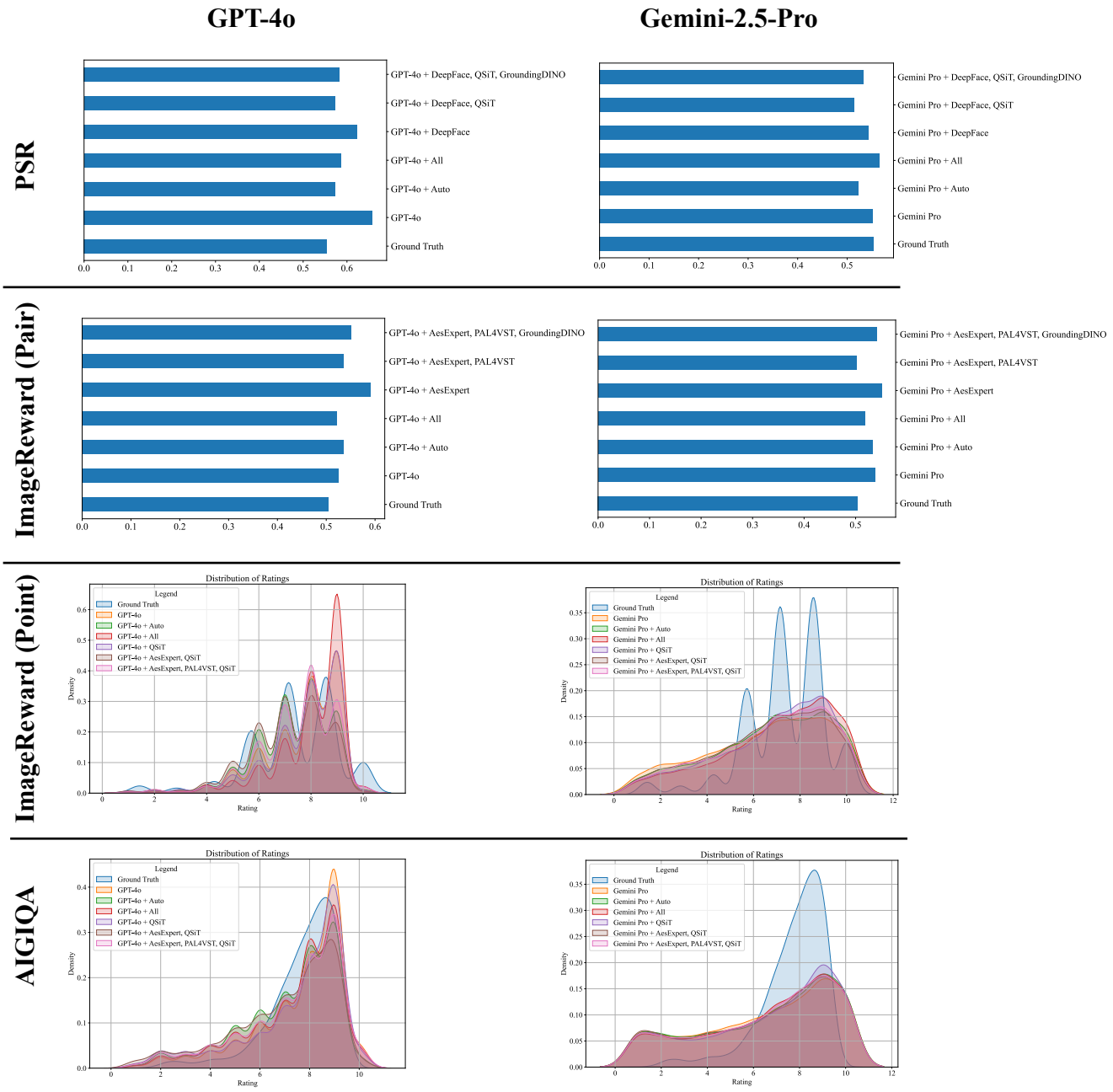

Figure 5. Judgment distributions of base VLMs and AUTORATER variants in different evaluation scenarios. Overall, AUTORATER variants have similar score distributions as their base VLMs. It indicates that the integration of specialists does not drastically alter the inherent scoring tendencies of the VLMs.

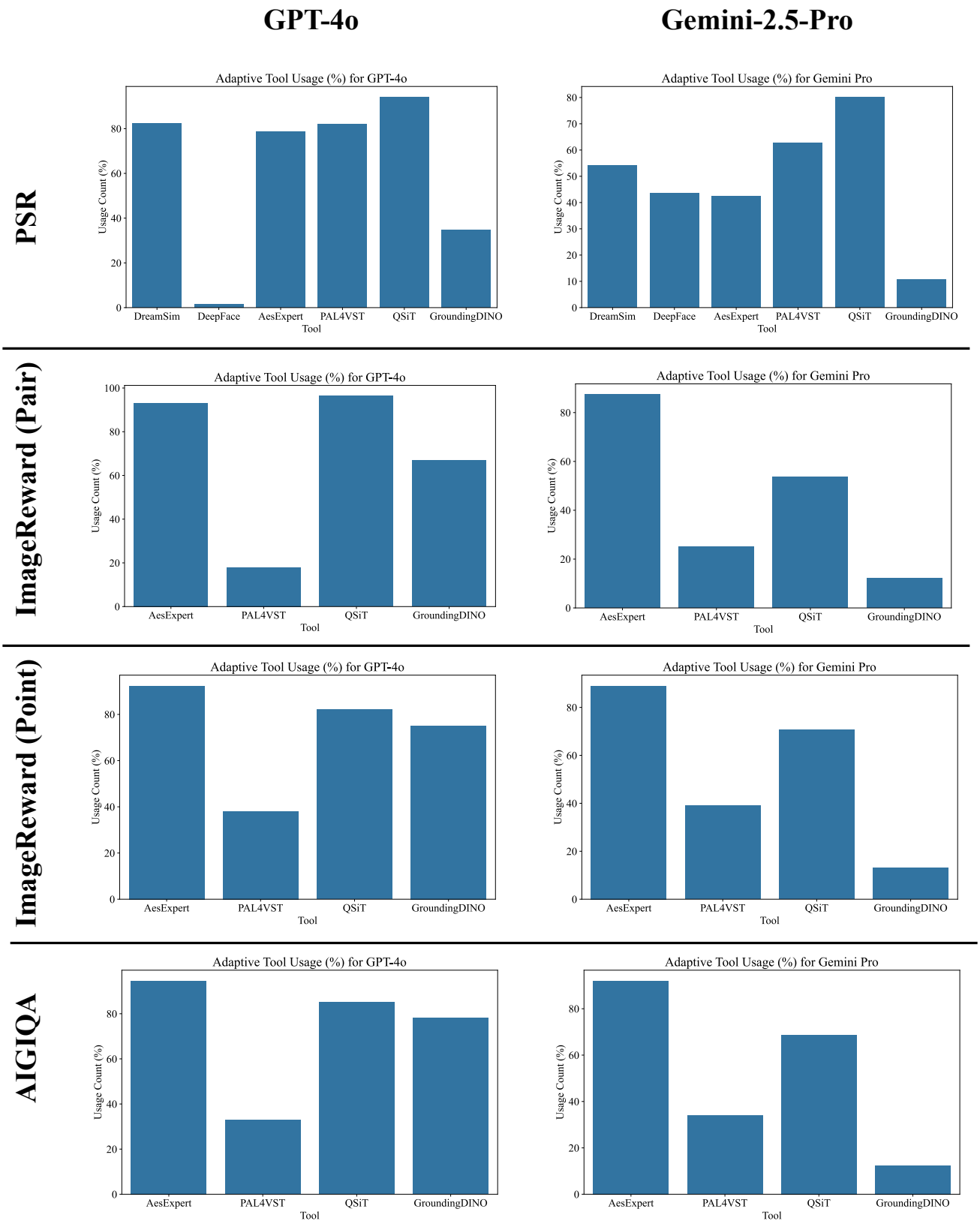

Figure 6. Statistics of tool selection results of AUTORATER with Auto tool selection mechanism in different evaluation scenarios. VLMs usually choose AesExpert and Q-SiT more frequently than other specialists across different evaluation scenarios. It suggests that aesthetic description from AesExpert and overall quality score from Q-SiT are generally useful for VLM judges to make decisions.