

Supplementary Material

A1. Additional details

A1.1. Prompts used to generated datasets

Figure A1. Long caption generation with GPT-4V

Please provide a detailed description of the image, ensuring that no details are omitted. Describe every element you observe within the image to provide a comprehensive account of its contents. Don't be lazy and it is important to get everything well done.

Figure A2. image-to-JSON data generation

First, provide a detailed description of the image, including every small detail possible. Next, create ten multiple-choice questions based on the content of the image. Each question should test the understanding of the image's content. Follow this JSON format: { "description": "Full Image Description", "short_description": "Short Image Description", "dialogue": ["Any visible dialogue text as a json list"], "on_screen_subtitle": "any subtitle on the image or n/a", "minimap_details": "Information from the minimap", "inventory_display": "Information about the player's inventory", "score_or_progress": "Details about scores or progress indicators", "NPC_status": "Information about NPCs", "event_indicators": "Indicators of any special events", "interaction_prompts": "Visible prompts for player interactions", "game_mode": "Current game mode or context", "HUD_description": "description of the game HUD or n/a if there is no HUD", "on_screen_watermark": "any watermark on the image or n/a", "summary_of_ui_values": "summary of the UI values as json or empty json if there is no UI", "scene_description": "A high-level overview of the entire scene", "character_list": [{ "name": "Character Name", "appearance": "Description of appearance", "clothing": "Description of clothing", "facial_expression": "Description of facial expression" }], "object_list": ["Object 1", "Object 2", ...], "texture_details": "a json list of object name and texture patterns that they have", "lighting_details": "Specific information about the light sources and shadows in the scene", "color_palette": ["hexadecimal color code", "hexadecimal color code", ...], "weather_conditions": "Description of any weather effects present, or say cannot be determined", "environmental_effects": "Description of any environmental effects like fog, rain, fire, etc.", "animation_states": "Descriptions of any static poses or actions implied by character positions", "error_log": "Any noticeable glitches or anomalies in the image", "glitches": "any glitch or buggy aspect of the image or none if there is nothing", "player_status": { "health": "Player's health value", "equipment": "Player's equipment details", "other_status": "Other status indicators" } }

Figure A3. LLama-3-based data generation

Using the image description provided below, create 10 questions and their corresponding answers that pertain exclusively to the details given in the description. Format your response using JSON.

Image description: <image description here >

Ensure your questions are relevant and directly related to the image description. For example, do not ask about elements not explicitly mentioned in the description.

Figure A4. GPT-4o-based data generation

First, provide a detailed description of the image, including every small detail possible. Next, create 10 questions answers based on the content of the image. Each question should test the understanding of the image's content.

Table A1. Description of entries in the JSON structure

Key	Description
Description	Detailed description of the image
Short description	Concise description of the image
Dialogue	A (JSON) list containing any visible dialogue text
On-Screen subtitle	Subtitles displayed on the image
Inventory display	Details of the player's inventory visible on the image
HUD description	Description of the game's Head-Up Display (HUD)
Scene description	High-level overview of the entire scene
NPC status	High level information about non-playable characters (NPCs)
Character list	List of characters, including their appearances, clothing, and facial expressions
Animation states	Descriptions of static poses or actions suggested by character positions
Object list	A (JSON) list containing all the visible objects in the scene
Texture details	A (JSON) list detailing object names and their texture patterns
Lighting details	Specific information about the light sources and shadows in the scene
Weather conditions	Description of any weather effects present, or state if they cannot be determined
Environmental effects	Description of environmental effects such as fog, rain, or fire
Player status	Player's health, equipment details, and other status indicators

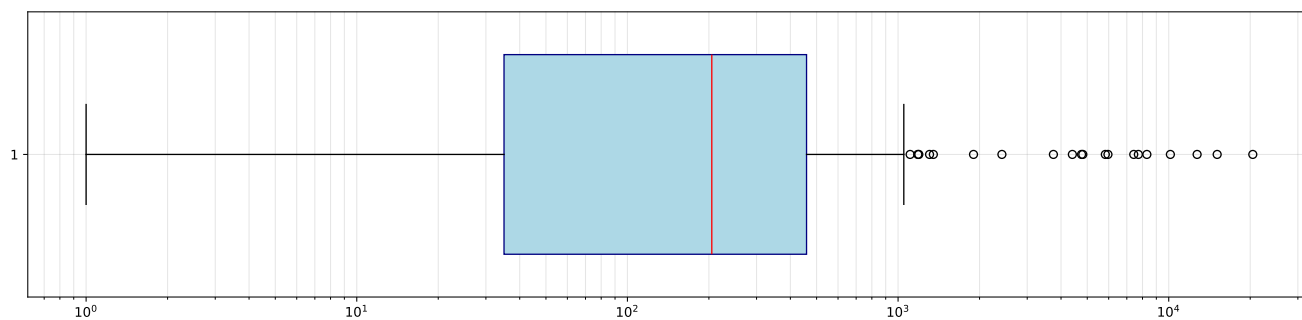


Figure A5. Image distribution across games, with a median of 205 unique images per game.

A2. Additional results

In this section, we provide complementary results for the experiments conducted in the main text.



Figure A6. Breakdown of improvement (percentage points) compared to the basemodel Bunny, after fine-tuning it on different mixture strategies by different question categories. As we increase the size of the dataset, all categories show improvement, with *Anomalies and Glitches* showing the greatest gains.

Table A2. Comparison of VIDEOGAMEBUNNY and LLaVA-1.6-34b

Category	VIDEOGAMEBUNNY	LLaVA-1.6-34b
Action Understanding	84.6 $\uparrow(+2.9)$	81.7
Anomalies and Glitches	82.1 $\uparrow(+16.6)$	65.5
Character Analysis	84.9 $\uparrow(+0.6)$	84.3
Common Sense Reasoning	89.1 $\downarrow(-2.8)$	91.9
Gameplay Mechanics	80.6 $\downarrow(-5.8)$	86.4
HUD and UI	84.1 $\uparrow(+3.0)$	81.1
Miscellaneous	86.2 $\uparrow(+1.7)$	84.5
Scene Understanding	92.4 $\uparrow(+0.2)$	92.2
Small Details	80.1 $\downarrow(-0.2)$	80.3
Spatial Reasoning	78.3 $\uparrow(+1.0)$	77.3

A3. Qualitative samples

In this section, we provide qualitative results from VIDEOGAMEBUNNY.

Qualitative results where VIDEOGAMEBUNNY provides a detailed description of the image.



Sullivan: Books, globe. This Want guy looks like a scholar.

Question: Briefly describe the image.

VIDEOGAMEBUNNY: The image is a screenshot from a video game, showing a character standing in front of a large, ornate painting. The painting depicts a man with long hair and a beard, wearing a period outfit with a high collar and a sash. The character is facing the painting, with their back to the viewer, and appears to be observing it. The room has a dimly lit, old-fashioned ambiance, with a sconce on the wall and a large, framed painting on the wall behind the character.

Figure A7. VIDEOGAMEBUNNY provides a detailed description of the image

Qualitative results where VIDEOGAMEBUNNY provides a detailed description of the image.



Question: Briefly describe the image.

VIDEOGAMEBUNNY: The image is a still from a video game, depicting two characters in a dimly lit wooden structure. The character on the left is a man with a stern expression, wearing a dark robe with a patterned design. He is facing the other character, who is partially obscured and has their back to the viewer. The second character appears to be wearing a similar robe. The scene is set in a rustic environment with wooden walls and a thatched roof.

Figure A8. VIDEOGAMEBUNNY provides a detailed description of the image

Qualitative results where VIDEOGAMEBUNNY provides a correct answer with minor inaccuracies



Question: Describe the characters in the image

VIDEOGAMEBUNNY: The image shows two characters on horseback. The character in the foreground is holding a lantern, and the character in the background *is following them*. Both characters are wearing hats.

Figure A9. VIDEOGAMEBUNNY provides a detailed description of the image with minor inaccuracies. Details that are incorrect are highlighted and emphasized in *red*.

Qualitative results where VIDEOGAMEBUNNY provides a correct answer with some inaccuracies



Question: Describe the characters in the image

VIDEOGAMEBUNNY: The image shows Batman, who is in the center of the frame, flying through the air. He is wearing his iconic black suit *with a yellow utility belt*. *Below him, there is a character named Penguin, who is wearing a tuxedo and a bowler hat.*

Figure A10. VIDEOGAMEBUNNY hallucinates extra characters and a Batman utility belt in the image, which are not actually present. Details that are incorrect are highlighted and emphasized in *red*.

Qualitative results where VIDEOGAMEBUNNY provide a **wrong** answer for a query



Question: What is written on the image?

VIDEOGAMEBUNNY: The text on the image reads: 'Tch... Tch... Tch... You way around it, I guess. Tch... Tch... Tch... I'll tell you, without stepping this guy away...'

Figure A11. A sample showing that VIDEOGAMEBUNNY struggles to convert the image to text.

Qualitative results where VIDEOGAMEBUNNY provide a **wrong** answer for a query



Question: What is written on the left?

VIDEOGAMEBUNNY: The word 'RELOAD' is written on the left side of the gun.

Figure A12. A sample showing that VIDEOGAMEBUNNY fails to provide a correct answer. While the word *RELOAD* is written in the center of the image, the query asks, *What is written on the left?* which should be answered with *nothing*.

A4. Sample image-to-JSON conversion

Sample image-to-JSON conversion done by VIDEOGAMEBUNNY



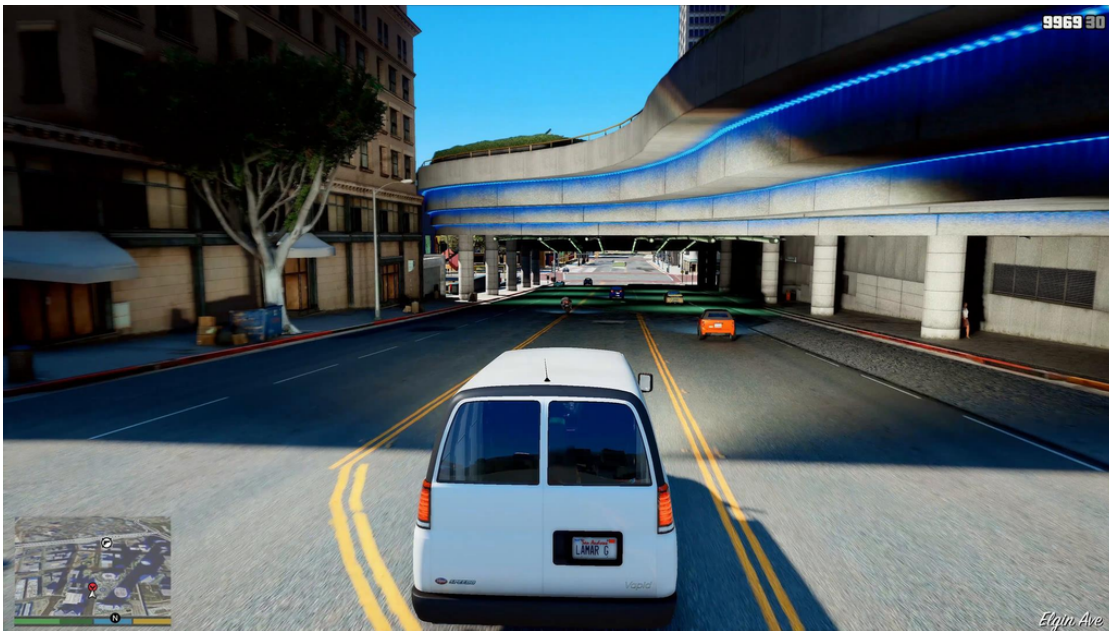
```
1 {
2   "description": "The image shows a first-person perspective of a video game, likely a survival horror title. The player is holding a large,
   → futuristic-looking rifle with a circular scope. The rifle has a white and black camouflage pattern. The player's left hand is visible, adorned with a
   → tattoo of a skull and what appears to be a snake. The player is standing in a dark, dilapidated wooden structure, possibly a barn or shed. The walls
   → are made of weathered wooden planks, and the floor is made of wooden planks as well. There is a doorway on the left side of the image, leading to a
   → darker area. A small, white silhouette of a person holding a rifle is visible on the left side of the screen, indicating the player's position on a
   → minimap. The minimap shows the player's location as a white dot within a circular area. The player's health bar is visible at the bottom of the
   → screen, represented by a green bar with a white outline. The health bar is nearly full. The player's ammunition count is displayed in the bottom
   → right corner of the screen, showing '0' bullets remaining in the current magazine and '41' bullets in reserve. A small icon of a skull with a knife
   → is visible next to the ammunition count, possibly indicating a special ability or weapon. The overall atmosphere of the image is dark, gritty, and
   → suspenseful, typical of the survival horror genre.",
3   "dialogue": [],
4   "on_screen_subtitle": "n/a",
5   "minimap_details": "The minimap shows the player's location as a white dot within a circular area. The minimap is located in the bottom left corner of
   → the screen.",
6   "inventory_display": "The image does not show the player's full inventory.",
7   "score_or_progress": "The image does not display any score or progress indicators.",
8   "NPC_status": "The image does not provide any information about NPCs.",
9   "event_indicators": "The image does not display any specific event indicators.",
10  "interaction_prompts": "The image does not show any interaction prompts.",
11  "game_mode": "The image suggests the game is in a single-player campaign mode, based on the presence of a minimap and health bar.",
12  "HUD_description": "The HUD consists of a circular minimap in the bottom left corner, a circular health bar in the bottom left corner, and an ammunition
   → counter in the bottom right corner. The ammunition counter displays the current magazine capacity, the remaining bullets in the magazine, and the
   → total reserve ammunition.",
13  "on_screen_watermark": "MKIceAndFire",
14  "summary_of_ui_values": {
15    "health": "Nearly full",
16    "ammunition": "0/41"
17  },
18  "scene_description": "The scene takes place inside a dilapidated wooden structure, likely a barn or shed. The structure is dark and appears to be
   → abandoned. The walls and floor are made of weathered wooden planks. There is a doorway on the left side of the image, leading to a darker area.",
19  "character_list": [
20    {
21      "name": "Player",
22      "appearance": "Cannot be fully determined, but the player's left hand is visible, showing a tattoo of a skull and a snake.",
23      "clothing": "Cannot be fully determined",
24      "facial_expression": "Cannot be determined"
25    }
26  ],
27  "object_list": [
28    "Rifle",
29    "Scope",
30    "Wooden planks (walls and floor)",
31    "Doorway"
32  ],
33  "texture_details": [
34    {
35      "object": "Rifle",
36      "texture": "White and black camouflage pattern"
37    },
38    {
39      "object": "Wooden planks",
```

```

40     "texture": "Weathered, with visible grain and knots"
41   }
42 },
43 "lighting_details": "The scene is dimly lit, with the primary light source appearing to be coming from the doorway on the left. This creates a sense of
→ depth and shadows within the structure.",
44 "color_palette": [
45   "#181818",
46   "#303030",
47   "#505050",
48   "#FFFFFF"
49 ],
50 "weather_conditions": "Cannot be determined",
51 "environmental_effects": "None",
52 "animation_states": "The player is holding the rifle in a ready position, suggesting a state of alertness and preparedness for combat.",
53 "error_log": "None",
54 "glitches": "None",
55 "player_status": {
56   "health": "Nearly full",
57   "equipment": "The player is equipped with a large, futuristic-looking rifle with a circular scope.",
58   "other_status": "The player has 0 bullets remaining in the current magazine and 41 bullets in reserve."
59 }
60 )

```

Sample image-to-JSON conversion done by VIDEOGAMEBUNNY



```

1 {
2   "description": "The image shows a third-person perspective of a video game, likely Grand Theft Auto V, based on the HUD elements and graphics style. The
→ player is driving a white van on a multi-lane road. The van is positioned slightly to the right of the center of the road, heading towards the
→ viewer. The road is a typical city street with sidewalks, buildings, and other vehicles. The time of day appears to be midday with clear skies. The
→ HUD elements visible include a minimap in the bottom left corner, a money counter in the top right corner, and a small portion of the player's
→ health bar in the bottom right corner. The minimap shows the player's location on a simplified map of the surrounding area. The money counter
→ displays the amount of in-game currency the player has, which is 9969. The health bar is partially obscured, but it appears to be mostly full.",
3   "dialogue": [],
4   "on_screen_subtitle": "n/a",
5   "minimap_details": "The minimap shows the player's location on a simplified map of the surrounding area. The player is represented by a white arrow
→ pointing upwards, indicating their direction. The map shows roads, buildings, and a body of water, likely a river or lake.",
6   "inventory_display": "Not visible in the image.",
7   "score_or_progress": "The player's current in-game money is displayed as 9969 in the top right corner.",
8   "NPC_status": "Not applicable, as there are no visible NPCs in the image.",
9   "event_indicators": "None visible in the image.",
10  "interaction_prompts": "None visible in the image.",
11  "game_mode": "Cannot be determined from the image.",
12  "HUD_description": "The HUD consists of a minimap in the bottom left corner, a money counter in the top right corner, and a small portion of the
→ player's health bar in the bottom right corner.",
13  "on_screen_watermark": "The watermark \"Elegy Ave\" is visible in the bottom right corner.",
14  "summary_of_ui_values": {
15    "money": "9969"
16  },
17  "scene_description": "The scene depicts a typical urban environment with buildings, roads, and vehicles. The player is driving a van on a multi-lane
→ road, heading towards a large, modern-looking building with a blue light display.",
18  "character_list": [],
19  "object_list": [
20    "Van",

```

```

21   "Buildings",
22   "Road",
23   "Sidewalk",
24   "Trees",
25   "Vehicles",
26   "Streetlights",
27   "Blue Light Display"
28 ],
29 "texture_details": [
30 {
31   "object": "Buildings",
32   "texture": "Brick, concrete, glass"
33 },
34 {
35   "object": "Road",
36   "texture": "Asphalt"
37 },
38 {
39   "object": "Sidewalk",
40   "texture": "Concrete"
41 }
42 ],
43 "lighting_details": "The scene is brightly lit, suggesting midday sun. The lighting casts soft shadows from the buildings and trees onto the road.",
44 "color_palette": [
45   "#87CEEB",
46   "#FFFFFF",
47   "#A9A9A9",
48   "#696969",
49   "#808080"
50 ],
51 "weather_conditions": "Clear and sunny.",
52 "environmental_effects": "None visible in the image.",
53 "animation_states": "The van's wheels suggest it is in motion, and the other vehicles on the road imply traffic.",
54 "error_log": "None visible in the image.",
55 "glitches": "None visible in the image.",
56 "player_status": {
57   "health": "Partially obscured, but appears to be mostly full.",
58   "equipment": "Not visible in the image.",
59   "other_status": "The player has 9969 in-game currency."
60 }
61 }

```

Sample image-to-JSON conversion done by VIDEOGAMEBUNNY



```

1 {
2   "description": "The image captures a scene from the video game Red Dead Redemption 2, showcasing the protagonist, Arthur Morgan, riding his horse, Black Jack, along a dirt path. The path leads towards a small, industrial settlement built on a cliffside. The settlement features wooden buildings with corrugated metal roofs, characteristic of the late 19th-century American frontier. A plume of black smoke rises from a chimney in the background, suggesting industrial activity. The surrounding environment is a mix of rocky terrain and lush vegetation, with a dense forest visible in the distance. The sky is overcast, with a hazy, diffused light suggesting either early morning or late afternoon. The overall atmosphere is one of quiet anticipation, as if Arthur is about to enter a new chapter in his journey.",
3   "dialogue": [],
4   "on_screen_subtitle": "n/a",

```

```
5  "minimap_details": "n/a",
6  "inventory_display": "n/a",
7  "score_or_progress": "n/a",
8  "npc_status": "n/a",
9  "event_indicators": "n/a",
10 "interaction_prompts": "n/a",
11 "game_mode": "n/a",
12 "HUD_description": "n/a",
13 "on_screen_watermark": "n/a",
14 "summary_of_ui_values": {},
15 "scene_description": "Arthur Morgan riding his horse Black Jack towards a small industrial settlement on a cliffside.",
16 "character_list": [
17   {
18     "name": "Arthur Morgan",
19     "appearance": "A rugged man with a beard and mustache, wearing a dark hat and a dark coat.",
20     "clothing": "Dark hat, dark coat, dark pants, and boots.",
21     "facial_expression": "Cannot be determined as his back is to the camera."
22   },
23 ],
24 "object_list": [
25   "Horse",
26   "Dirt path",
27   "Wooden buildings",
28   "Corrugated metal roofs",
29   "Smoke plume",
30   "Chimney",
31   "Forest",
32   "Rocks",
33   "Vegetation",
34   "Trees"
35 ],
36 "texture_details": [
37   {
38     "object": "Horse",
39     "texture": "Short, dark brown fur"
40   },
41   {
42     "object": "Dirt path",
43     "texture": "Uneven, dusty, with patches of grass"
44   },
45   {
46     "object": "Wooden buildings",
47     "texture": "Weathered wood planks"
48   },
49   {
50     "object": "Corrugated metal roofs",
51     "texture": "Rusted and dented metal sheets"
52   },
53   {
54     "object": "Rocks",
55     "texture": "Rough, jagged, with moss and lichen"
56   },
57   {
58     "object": "Vegetation",
59     "texture": "A mix of grass, shrubs, and trees"
60   }
61 ],
62 "lighting_details": "The scene is lit by diffused sunlight, suggesting an overcast sky. The light is soft and even, with no strong shadows.",
63 "color_palette": [
64   "#A0A0A0",
65   "#808080",
66   "#606060",
67   "#404040",
68   "#202020",
69   "#008000",
70   "#006400",
71   "#004000"
72 ],
73 "weather_conditions": "Overcast",
74 "environmental_effects": "Smoke plume rising from a chimney",
75 "animation_states": "Arthur Morgan is sitting upright on his horse, suggesting a calm and controlled ride.",
76 "error_log": "None",
77 "glitches": "None",
78 "player_status": {
79   "health": "n/a",
80   "equipment": "n/a",
81   "other_status": "n/a"
82 }
83 }
```

Sample image-to-JSON conversion done by VIDEOGAMEBUNNY



```
1 {
2   "description": "The image showcases a close-up of a man in a World War II-era military uniform. He is positioned slightly to the right of the center,
   → his gaze directed slightly upwards and to the left. His expression is serious, bordering on somber. He has short, dark hair and a prominent nose.
   → The man's uniform is a muted green, with a noticeable dirt smudge on his left cheek. A dark green helmet with a netting pattern covers his head,
   → partially obscuring his hair. The helmet's chin strap is visible, fastened with a metal buckle. A thick, dark green strap crosses his chest, likely
   → part of a backpack or equipment harness. The background is a blurred depiction of a battlefield, with other soldiers in similar uniforms visible in
   → the distance. The overall color palette is muted and desaturated, reflecting the grim atmosphere of war.",
3   "dialogue": [],
4   "on_screen_subtitle": "n/a",
5   "minimap_details": "n/a",
6   "inventory_display": "n/a",
7   "score_or_progress": "n/a",
8   "NPC_status": "n/a",
9   "event_indicators": "n/a",
10  "interaction_prompts": "n/a",
11  "game_mode": "n/a",
12  "HUD_description": "n/a",
13  "on_screen_watermark": "n/a",
14  "summary_of_ui_values": {},
15  "scene_description": "The scene depicts a soldier in the midst of a battlefield, likely during World War II, given the uniforms and equipment.",
16  "character_list": [
17    {
18      "name": "Unknown",
19      "appearance": "The man appears to be in his late twenties to early thirties. He has a rugged appearance with short, dark hair and a prominent nose.",
20      "clothing": "He is wearing a World War II-era military uniform in a muted green color. A dark green helmet with a netting pattern covers his head,
      → and a thick, dark green strap crosses his chest, likely part of a backpack or equipment harness.",
21      "facial_expression": "His expression is serious and slightly somber, suggesting he is focused on the situation at hand."
22    }
23  ],
24  "object_list": [
25    "Military uniform",
26    "Helmet",
27    "Backpack strap",
28    "Backpack (partially visible)",
29    "Soldiers (in the background)"
30  ],
31  "texture_details": [
32    {
33      "object": "Helmet",
34      "texture": "Netting pattern"
35    },
36    {
37      "object": "Uniform",
38      "texture": "Camouflage pattern"
39    }
40  ],
41  "lighting_details": "The lighting is soft and diffused, likely due to overcast skies or fog. It casts subtle shadows on the man's face, accentuating his
   → features.",
42  "color_palette": [
43    "#5C6B5A",
44    "#3D4A3B",
45    "#7A8A76",
46    "#B2B4A9"
47  ],
48  "weather_conditions": "The weather appears to be overcast or foggy, as evidenced by the soft lighting and lack of harsh shadows.",
```

```

49 "environmental_effects": "The background suggests a battlefield, with smoke or dust particles visible in the air.",
50 "animation_states": "The man appears to be in a still pose, likely captured in a moment of pause or reflection during the chaos of battle.",
51 "error_log": "None",
52 "glitches": "None",
53 "player_status": {
54   "health": "n/a",
55   "equipment": "n/a",
56   "other_status": "n/a"
57 }
58 }

```

Sample image-to-JSON conversion done by VIDEOGAMEBUNNY



```

1 {
2   "description": "The image shows a male character in a video game, seemingly in the middle of a climbing or jumping action. He is positioned on a stone
  → platform within a dimly lit, ancient-looking structure. The platform is part of a larger architectural complex, with visible stone walls and a
  → barred gate in the background. The character is dressed in a sleeveless top and pants, suggesting a warrior or adventurer archetype. He is equipped
  → with a sword sheathed on his back and appears to be in motion, with his body angled towards the platform's edge. The overall atmosphere is one of
  → adventure and exploration, with the character's pose hinting at a moment of action or transition within the game.",
3   "dialogue": [],
4   "on_screen_subtitle": "n/a",
5   "minimap_details": "Cannot be determined",
6   "inventory_display": "Not visible",
7   "score_or_progress": "Not visible",
8   "NPC_status": "No NPCs present",
9   "event_indicators": "None visible",
10  "interaction_prompts": "None visible",
11  "game_mode": "Cannot be determined",
12  "HUD_description": "n/a",
13  "on_screen_watermark": "SHIRRAKO in the bottom right corner",
14  "summary_of_ui_values": {},
15  "scene_description": "The scene is set within a large, ancient-looking structure, possibly a temple or palace. The architecture suggests a Middle
  → Eastern or Mediterranean influence. The environment is dimly lit, with the primary light source appearing to be natural light filtering in from an
  → unseen opening above.",
16  "character_list": [
17    {
18      "name": "Unknown",
19      "appearance": "Male, athletic build",
20      "clothing": "Sleeveless top, pants, possibly leather or fabric",
21      "facial_expression": "Cannot be determined"
22    }
23  ],
24  "object_list": [
25    "Stone platform",
26    "Stone walls",
27    "Barred gate",
28    "Sword",
29    "Pottery jug"
30  ],
31  "texture_details": [
32    {
33      "object": "Stone walls",
34      "texture": "Rough, uneven stone with visible cracks and weathering"

```

```
35     },
36     {
37         "object": "Stone platform",
38         "texture": "Similar to the walls, but with a slightly smoother appearance"
39     }
40 ],
41 "lighting_details": "The scene is dimly lit, with the primary light source appearing to be natural light coming from an unseen opening above. This
→ creates a sense of depth and highlights the textures of the stone surfaces.",
42 "color_palette": [
43     "#5C4A3A",
44     "#3B322A",
45     "#7A6D5B",
46     "#A1887F"
47 ],
48 "weather_conditions": "Cannot be determined",
49 "environmental_effects": "None visible",
50 "animation_states": "The character's pose suggests a dynamic action, either in the middle of a jump or a climbing maneuver.",
51 "error_log": "None visible",
52 "glitches": "None",
53 "player_status": {
54     "health": "Not visible",
55     "equipment": "Sword visible on back",
56     "other_status": "Not visible"
57 }
58 }
```