GPT と Qwen3-VL-4B の OCR 比較結果

使用モデル: GPT-5.2 Thinking, Qwen3-VL-4B-Instruct

表 1 に cut, action memo, dialogue, time の結果を示す. GPT と Qwen3-VL-4B の差はほとんどなく, “page1 row1 cut” でのみ差があった. また, “page2 row2 dialogue” ではどちらのモデルも “95+12” と出力しているが, 入力画像をみると正解は “05+12” であり, どちらも間違いとなった. 以下に載せている結果以外はすべてどちらも正解していた.

表 2 に picture の結果を示す. GPT-5.2 Thinking は出力があったのに対し, Qwen3-VL-4B はすべて ‘□’ を出力して読み取れるものがないという判定になった. GPT-5.2 Thinking は画像の左下に小さく書かれている文字を認識しているが, 元の画像をみると “HDTV 用” と書かれており, 間違えた出力となっている. また, 赤色の矢印および文字はどちらのモデルもうまく読み取れていない, もしくは間違えていたが, GPT-5.2 Thinking のほうが正解に近い結果となった.

表 1: cut, action memo, dialogue, time

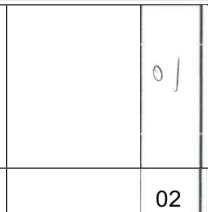
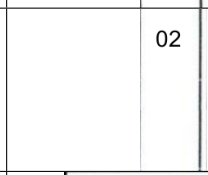
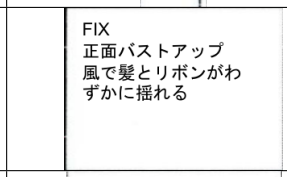
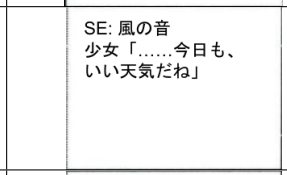
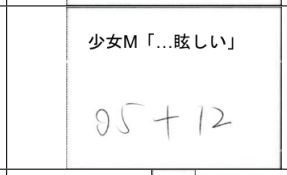
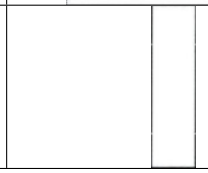
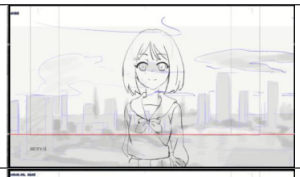
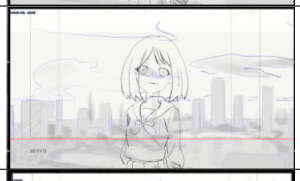
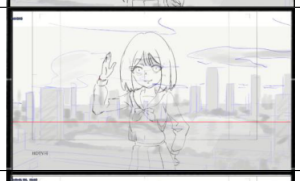
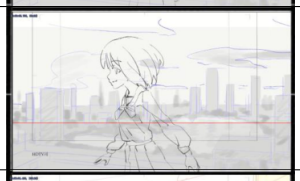
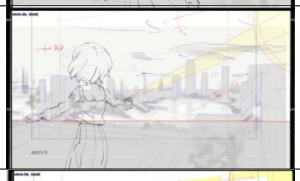
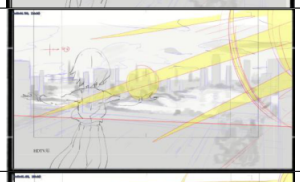
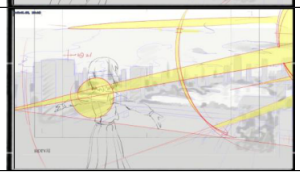
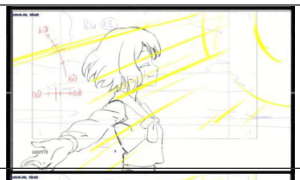
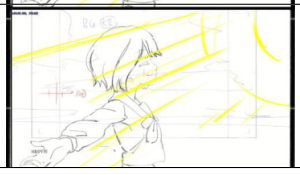
ID	入力画像	GPT-5.2 Thinking	Qwen3-VL-4B-Instruct
page1 row1 cut		01	O.L
page1 row3 cut		02	02
page1 row1 action memo		FIX 正面バストアップ 風で髪とリボンが わずかに揺れる	FIX 正面バストアップ 風で髪とリボンが わずかに揺れる
page1 row1 dialogue		SE: 風の音 少女「……今日も、 いい天気だね」	SE: 風の音 少女「……今日も、 いい天気だね」
page2 row2 dialogue		少女 M「…眩しい」 95 + 12	少女 M「…眩しい」 95 + 12
page1 row1 time		□	□

表 2: picture

ID	入力画像	GPT-5.2 Thinking	Qwen3-VL-4B-Instruct
page1 row1 picture		MIDTV1	□
page1 row2 picture		MIDTV4	□
page1 row3 picture		HDTV4	□
page1 row4 picture		HDTV14	□
page1 row5 picture		HDTV	□
page2 row1 picture		CUT 0050 HDTV/1	□
page2 row2 picture		No. HDTV/8	□
page2 row3 picture		No. 0000 BG ② B0 ② A0 HDTV/8	□
page2 row4 picture		A0 HDTV/8	□
page2 row5 picture		HDTV10	□