Here is the updated **PASS/REJECT Scoring & Evaluation System**, strictly aligned with the **Decisions**, **Gotchas**, and **Audit Rubric** defined in the Input Pack.

1. Hard Gates ("The Kill Switch")

*These checks run programmatically before detailed scoring. If **ANY** fail, the frame is immediately REJECTED (Score = 0). This enforces the "Pipeline Constraints" and prevents wasting compute on unusable assets.*

Gate Code	**Metric / Condition**	**Threshold**	**Rationale (Decisions & Gotchas)**
HG_DIM	width != anchor_w OR height != anchor_h	**Exact Match**	**Decision:** "Frame size must match anchor dimensions exactly." Trimming is an *export* step (TexturePacker); the source frame must remain full-size to preserve the (0.5, 1.0) pivot logic.
HG_BASE	baseline_error_px	**> 2.0 px**	**Gotcha:** "Baseline consistency matters most." While ±1px is the soft target, >2px creates visible "ice skating" that breaks game feel.
HG_ID	vision_id_score	**≥ 4** (1-5 scale)	**Rubric:** Score 4 ("Poor") or 5 ("Unrecognizable") is a hard failure. Identity is locked to the anchor; no redesigns allowed.
HG_ALPHA	alpha_integrity	**Opaque / Block**	**Constraint:** Background must be transparent. Detects failure to mask/matte the subject or "black box" backgrounds.
HG_SAFE	bbox_touching_border	**True**	**Gotcha:** If pixels touch the edge, TexturePacker's --trim-mode Trim +

			--extrude 1 can cause artifact bleeding or limb truncation.

2. Score Formula (Total ∈ [0, 100])

*The formula prioritizes **Stability** and **Identity** (Hygiene Factors) over Pose Perfection.*

$$ \text{Score} = 100 \times (w_{St} S_{St} + w_{Id} S_{Id} + w_{Pal} S_{Pal} + w_{Sty} S_{Sty}) $$

Weights (Prioritized by Project Constraints)

- **$w_{St}$ (Stability) = 0.35**: "Baseline/pivot consistency matters most."
- **$w_{Id}$ (Identity) = 0.30**: "Identity is locked to anchor."
- **$w_{Pal}$ (Palette) = 0.20**: "Enforce canonical colors." (Critical, but often fixable in post).
- **$w_{Sty}$ (Style/Struct) = 0.15**: "Style must match." Lower weight allows for valid animation deformation (squash/stretch).

Component Normalization

1. **Stability ($S_{St}$):** Exponential decay for pixel-perfect grounding.
    - $$ S_{St} = e^{-1.5 \times \text{baseline\_error\_px}} $$
    - *Impact:* 0px = **1.0** (Perfect) | 1px = **0.22** (Harsh Penalty) | 2px = **0.05** (Fail).
2. **Identity ($S_{Id}$):** Linear mapping of Vision Score.
    - $$ S_{Id} = 1.0 - \frac{\text{vision\_id\_score} - 1}{3.0} $$
    - *Note:* Input range is effectively 1–3 (since ≥4 is Gated).
3. **Palette ($S_{Pal}$):** Deviation from Spec.
    - $$ S_{Pal} = 1.0 - (\text{palette\_delta} \times 3) $$
    - *Correction:* Must use **Sanitized Hex List** (see Section 6) to avoid "Sean Tank" typo failures.
4. **Style ($S_{Sty}$):** Composite of Edge Map (Cleanliness) and SSIM (Structure).
    - $$ S_{Sty} = 0.6(\text{edge\_map\_sim}) + 0.4(\text{ssim}) $$
    - *Resolution:* SSIM is weighted low to allow for valid animation movement, resolving the "0.85 vs 0.95" conflict by treating it as a secondary check.

3. Thresholds & Status Table (Resolving SSIM Conflict)

*We resolve the contradictory SSIM thresholds (<0.75 vs >0.95) by using a **Tiered Soft Fail** system.*

Rank	**Score Range**	**SSIM Check**	**Status**	**Action Strategy**
Diamond	**92 - 100**	**> 0.90**	**PASS**	**Auto-Commit.** Pixel-perfect stability (0px) and high fidelity.
Gold	**80 - 91**	**> 0.85**	**PASS**	**Acceptable.** Likely has 1px jitter or minor alpha noise.
Silver	**65 - 79**	**< 0.85**	**SOFT FAIL**	**Conditional Retry.** If attempts < 3, **RETRY**. If attempts == 3, flag for **Manual Review** (Batch B).
Bronze	**0 - 64**	Any	**REJECT**	**Hard Fail.** Discard and trigger Retry Mapping.

4. Reject Reason Code Mapping

*When a frame fails, assign the **primary** reason code to guide the Agent.*

Code	**Trigger Condition**	**Description**
REJ_JITTER	baseline_error_px > 1.0	**Stability.** Feet do not align with anchor. "Ice skating" risk.
REJ_ID	vision_id_score ≥ 3	**Identity.** Face/Costume details lost. "Who is this?"
REJ_PAL	palette_delta > 0.15	**Color.** Hallucinated colors or lighting effects (gradients).

REJ_HALO	alpha_fringe > 0.05	**Alpha.** "Dirty" edges. 16BitFit requires clean transparency (Gotcha fix).
REJ_STYLE	edge_map_sim < 0.6	**Texture.** Blurry lines, painterly texture, or "AI slop" noise.
REJ_BROKEN	ssim < 0.6	**Pose.** Major anatomical failure (missing limbs/head).

5. Retry Mapping (The "Knob Ladder")

Mapped directly to the "Retry Ladder" steps in Section 4 of the Input Pack.

Reject Code	**Primary Action (Ladder Step)**	**Specific Agent Knob Tweak**
REJ_JITTER	**Step 4: Pose Rescue**	Increase **ControlNet (Pose)** weight (+0.15). Enforce bottom_center pivot. *Do not re-prompt.*
REJ_ID	**Step 3: Identity Rescue**	Increase **IP-Adapter/Reference** strength (+0.1). Decrease **Denoise** (-0.05).
REJ_PAL	**Step 6: Post-Process**	**Quantize:** Force-remap pixels to Anchor Palette (Sanitized). Check Prompt for "Color contamination."
REJ_HALO	**Step 2: Negative Prompt**	Add: "anti-aliasing, semi-transparent, halo, bloom, glow".

REJ_STYLE	**Step 2: Negative Prompt**	Add: "blur, painterly, 3d render, vector art". Ensure downscale_mode=NEAREST.
REJ_BROKEN	**Step 1: Reroll**	**Reroll Seed.** Structural failures are often random noise.

---

6. Tuning & Gotcha Resolutions (Auditable)

1. **The "Sean Tank" Correction (Spec Data Integrity)**
   - **Gotcha:** Spec hex #F2FOEF contains a typo ('O' instead of '0').
   - **Resolution:** The Scoring Function must run a **Palette Sanitizer** pre-check. If 'O' is detected in any hex string, the system must **sample the Anchor Sprite's center region** to resolve the true hex (likely #F2F0EF) *before* calculating palette_delta.
2. **Trim vs. Dimensions**
   - **Gotcha:** "Trimming can cause baseline drift."
   - **Resolution:** The QA system enforces **Full Frame Integrity** (Hard Gate HG_DIM). The generator must not trim. Trimming is strictly delegated to TexturePacker export using --trim-mode Trim.
3. **Phaser Pivot Logic**
   - **Gotcha:** "Phaser reads pivot... is NOT confirmed."
   - **Resolution:** Audit logs must record the calculated pivot (e.g., pivot_x: 0.5, pivot_y: 1.0). If REJ_JITTER rates are high, the Loader Script must be updated to enforce frame.customPivot = true explicitly.
4. **Suffix Mismatch**
   - **Gotcha:** .png suffix breaks animation generation.
   - **Resolution:** The scoring system's "Tie-Break" logic prefers filenames matching the TexturePacker convention (suffix: '').

Here is the updated **PASS/REJECT Scoring & Evaluation System (v2.0)**.

This specification integrates your **Decisions** (TexturePacker export, Identity Lock) and **Gotchas** (baseline jitter, alpha halos) into the scoring logic. It replaces subjective "vision scores" with **LPIPS** and **DINOv2**—industry-standard, computable metrics for perceptual quality and semantic identity.

1. Hard Gates ("The Kill Switch")

*These checks run programmatically **before** detailed scoring. If **ANY** fail, the frame is immediately REJECTED (Score = 0). This enforces pipeline constraints and prevents "garbage" from wasting compute.*

Gate Code	Metric / Condition	Threshold	Rationale (Decisions & Gotchas)
HG_DIM	width != anchor_w OR height != anchor_h	**Mismatch**	**Decision:** Frames must match Anchor dimensions *exactly* to preserve the (0.5, 1.0) pivot logic. Trimming is an *export-only* step.
HG_BASE	baseline_error_px	**> 2.0 px**	**Gotcha:** "Baseline consistency matters most." >2px creates visible "ice skating." (Soft target is 0–1px).
HG_ID	dinov2_similarity	**< 0.60**	**New Metric:** DINOv2 Cosine Similarity. < 0.60 implies a different character (wrong species/gender) or severe deformity.
HG_ALPHA	alpha_integrity	**Opaque**	**Constraint:** Background must be transparent. Detects failure to mask/matte or "black box" generation.
HG_SAFE	bbox_touching_border	**True**	**Gotcha:** If pixels touch the edge, TexturePacker's --trim-mode Trim + --extrude 1 causes artifact bleeding in the atlas.

2. Score Formula (Total ∈ [0, 100])

*The formula prioritizes **Stability** and **Identity** (Hygiene Factors) over Pose Perfection.*

$$ \text{Score} = 100 \times (w_{St} S_{St} + w_{Id} S_{Id} + w_{Pal} S_{Pal} + w_{Tex} S_{Tex}) $$

Weights & Normalization

1. **w_{St} Stability (35%)**: "Baseline consistency is King."

    - $$ S_{St} = e^{-1.5 \times \text{baseline_error_px}} $$
    - *Impact:* 0px = **1.0** (Perfect) | 1px = **0.22** (Harsh Penalty) | 2px = **0.05** (Fail).
2. **w_{Id} Identity (30%)**: "Identity is locked to anchor."
    - $$ S_{Id} = \text{clamp}\left(\frac{\text{dino_sim} - 0.60}{0.40}, 0, 1\right) $$
    - *Metric:* **DINOv2 (ViT-S/14)** Cosine Similarity.
    - *Why:* Robust to pose changes (unlike SSIM). Verifies "Is this Sean?" even if he is kicking.
3. **w_{Pal} Palette (20%)**: "Enforce canonical colors."
    - $$ S_{Pal} = 1.0 - (\text{palette_delta} \times 3.0) $$
    - *Correction:* System **MUST** run sanitize_palette() (see Section 6) to fix the #F2FOEF typo *before* scoring.
4. **w_{Tex} Style/Texture (15%)**: "Art style coherence."
    - $$ S_{Tex} = 1.0 - \text{clamp}\left(\frac{\text{LPIPS}}{0.3}, 0, 1\right) $$
    - *Metric:* **LPIPS (AlexNet)**.
    - *Why:* Detects "blur", "AI slop", and "painterly" artifacts that SSIM misses. Lower LPIPS (<0.1) is better.[1]

3. Thresholds & Status Table

Resolves the SSIM conflict by demoting it to a secondary check and relying on LPIPS/DINO for quality.

Rank	**Score Range**	**Logic**	**Status**	**Action Strategy**
Diamond	**92 - 100**	baseline_err == 0	**PASS**	**Auto-Commit.** Pixel-perfect stability and high fidelity.
Gold	**80 - 91**	LPIPS < 0.18	**PASS**	**Acceptable.** Likely 1px jitter or minor color noise (fixable).
Silver	**65 - 79**	DINO > 0.75	**SOFT FAIL**	**Conditional Retry.** If attempts < 3, **RETRY**. Else, **FLAG** for Human Batch B.

Bronze	**0 - 64**	Any	**REJECT**	**Hard Fail.** Discard and trigger Retry Mapping.

4. Reject Reason Code Mapping

Assigns the failure code based on the lowest component score to guide the Agent.

Code	Trigger	Description
REJ_JITTER	baseline > 1.0	**Stability.** Feet do not align with anchor. "Ice skating" risk.
REJ_ID	dino < 0.75	**Identity.** Character unrecognizable / wrong outfit features.
REJ_BLUR	lpips > 0.25	**Style.** Image is "mushy", painterly, or lacks pixel definition.
REJ_PAL	pal_delta > 0.1	**Color.** Hallucinated colors, gradients, or wrong team colors.
REJ_HALO	alpha_fringe > 5%	**Alpha.** Dirty edges/halos. (Gotcha: Requires ReduceBorderArtifacts fix).
REJ_BROKEN	ssim < 0.6	**Pose.** Major anatomical failure (missing limbs).

5. Retry Mapping ("The Knob Ladder")

Directly maps Reject Codes to Agent Actions.

Reject Code	**Primary Action (Ladder Step)**	**Specific Agent Knob Tweak**
REJ_JITTER	**Step 4: Pose Rescue**	Increase **ControlNet (Pose)** weight (+0.15). Enforce bottom_center pivot. *Do not re-prompt.*
REJ_ID	**Step 3: Identity Rescue**	Increase **IP-Adapter** weight (+0.1). Decrease **CFG Scale** (high CFG "burns" features).
REJ_BLUR	**Step 2: Negative Prompt**	Add: "blur, anti-aliasing, painterly, 3d render". Force downscale_mode=NEAREST.
REJ_PAL	**Step 6: Post-Process**	**Quantize:** Force-remap pixels to Anchor Palette (Sanitized). Check Prompt for color contamination.
REJ_HALO	**Step 6: Cleanup**	Run a generic **Erosion** filter (1px) on the alpha channel or re-matte.
REJ_BROKEN	**Step 1: Reroll**	**New Seed.** Structural failures are often stochastic noise.

6. Implementation & Logic Updates

1. **DINOv2 Implementation:**
   - Use torch.hub.load('facebookresearch/dinov2', 'dinov2_vits14') (Small is fast/sufficient).
   - **Metric:** Cosine Similarity of the [CLS] token.
2. **LPIPS Transparency Hack (Crucial):**
   - LPIPS expects RGB and fails on alpha.[2] **Do not** pass raw RGBA.
   - **Action:** Composite both Anchor and Candidate onto a **neutral grey (#808080)** background before scoring. This ensures the silhouette shape is evaluated as part of the style.
3. **Palette Sanitizer (Gotcha Resolution):**

    - The system must run a sanitize_spec() pre-step. If hex #F2FOEF (letter O) is detected, it must auto-correct to #F2F0EF (digit 0) by sampling the Anchor Sprite's center region, rather than hard-failing the batch.