Intelligenz Welt

Qwen3-VL-235B-A22B

KI-Modellportrait · Quellenstand 2026-10-31

QV

Monogramm-Fallback; kein sauber belegtes Logo-Asset (Wikimedia-Dateien mit Marken- und Lizenz-Risiko, siehe Integrationsnotiz) Monogramm-Darstellung, kein Logo-Asset übernommen.

Offenes Vision-Language-Flaggschiff der Qwen-Reihe von Alibaba Cloud: Mixture-of-Experts mit 235 Mrd. Parametern (22 Mrd. aktiv), als Instruct- und als Thinking-Edition veröffentlicht.

Quellen geprüft am 31. Oktober 2026 · Redaktion: Philipp Paulik

Entwickler: Alibaba Cloud →

DAS MODELL EINFACH ERKLÄRT

Offenes Multimodal-Flaggschiff mit Agenten-Fokus richtig einordnen.

Qwen3-VL-235B-A22B ist das größte Vision-Language-Modell der Qwen3-VL-Serie von Alibaba Cloud. Es verarbeitet Text, Bilder und Videos und erzeugt Text. Die Gewichte stehen unter Apache 2.0 in zwei Editionen bereit: Instruct und Thinking. In der Model-Studio-API stehen 131.072 Token Kontext und maximal 32.768 Ausgabe-Token.

Offizielle Dokumentation: Qwen3-VL-235B-A22B

Quellen geprüft: 31.10.2026. Herstellerangaben, kein eigener Modelltest.

Wofür eignet sich Qwen3-VL-235B-A22B?

Alibaba positioniert das Modell für visuelles Reasoning, Dokument- und Videoanalyse, OCR in 32 Sprachen und agentische Aufgaben wie das Bedienen von Oberflächen. Ob Instruct oder Thinking im eigenen Projekt passt, hängt von Aufgabenart, Latenz, Kosten und davon ab, ob eine sichtbare Denkphase gewünscht ist - nicht allein von der Parameterzahl.

Instruct und Thinking getrennt einordnen

Beide Editionen teilen Architektur und Größe, haben aber eigene Gewichte, eigene API-IDs und unterschiedliche Preise: Die Thinking-Ausgabe kostet in der API das 2,5-Fache der Instruct-Ausgabe. Thinking ist laut Hersteller für STEM- und Mathematik-Reasoning optimiert; Instruct antwortet direkt. Ein Wechsel der Edition ist ein Modellwechsel, kein Schalter in derselben Anfrage.

Offizielle Dokumentation: Qwen3-VL-235B-A22B

Agentische Fähigkeiten: Oberflächen bedienen

Alibaba beschreibt Qwen3-VL als Visual Agent, der PC- und Mobil-Oberflächen erkennt, Elemente versteht, Werkzeuge aufruft und Aufgaben abschließt; bei OS World nennt der Hersteller Spitzenwerte. Aus Bildern oder Videos kann das Modell Code erzeugen, etwa Draw.io, HTML, CSS oder JavaScript. Diese Angaben sind Herstellerzusagen und ersetzen keinen Test mit der eigenen Oberfläche.

Offizielle Dokumentation: Qwen3-VL-235B-A22B

Kontext: 256K nativ, 1M erweiterbar, API 128K

Die Gewichte unterstützen nativ 256K Token und lassen sich auf 1M erweitern - genug für Bücher oder stundenlange Videos. In einem Hersteller-Nadeltest erreichte das Modell 100 % Treffer bei 256K und 99,5 % bei 1M Token Video. Die gehostete API begrenzt Anfragen dagegen auf 131.072 Token Fenster. Wer lange Videos verarbeiten will, muss Zugangsweg und Budget getrennt prüfen.

Offizielle Dokumentation: Qwen3-VL-235B-A22B

OCR und Dokumentverständnis

Die OCR-Fähigkeit deckt laut Modellkarte 32 Sprachen ab und soll bei schwachem Licht, Unschärfe und schrägem Text robust bleiben; auch seltene und historische Zeichen sowie Fachbegriffe werden genannt. Für Dokumente beschreibt der Hersteller eine Struktur-Analyse mit Layout-Positionsangaben in einem Qwen-HTML-Format.

Offizielle Dokumentation: Qwen3-VL-235B-A22B

Architektur: MoE mit neuer Sehen-Pipeline

Das Modell ist ein Mixture-of-Experts mit 235 Mrd. Parametern, von denen pro Token 22 Mrd. aktiviert werden. Drei Änderungen nennt der Hersteller: Interleaved-MRoPE verteilt Zeit-, Höhen- und Breiteninformation über alle Frequenzen für besseres Langvideo-Verständnis; DeepStack speist ViT-Merkmale mehrerer Ebenen in mehrere LLM-Schichten ein; eine Text-Zeitstempel-Ausrichtung verankert Ereignisse in Videos sekundengenau.

Offizielle Dokumentation: Qwen3-VL-235B-A22B

Selbst hosten oder API nutzen

Die Gewichte stehen unter Apache 2.0 bei Hugging Face und ModelScope; der Einsatz setzt transformers ab Version 4.57.0 voraus, flash_attention_2 wird empfohlen. Allein die BF16-Gewichte entsprechen rechnerisch rund 470 GB Speicher, dazu kommt Aktivierungsspeicher - Selbsthosting ist ein Multi-GPU-Projekt. Die API bei Model Studio ist der einfachere Einstieg, unterliegt aber Ratenlimits von 60 Anfragen und 100.000 Token pro Minute.

Offizielle Dokumentation: Qwen3-VL-235B-A22B

Kontextfenster ist nicht Ausgabelimit.

131.072

Token Kontextfenster

32.768

maximale Ausgabe-Token, synchron



Zwei unterschiedliche Grenzen auf gemeinsamer Nullskala, nicht addieren. Offizielle Dokumentation: Qwen3-VL-235B-A22B

Was kostet Qwen3-VL-235B-A22B über die API?

Standard-Eingabe

2 USD

je 1 Mio. Texttoken · Basisrate

Standard-Ausgabe

8 USD

je 1 Mio. Texttoken · Basisrate

Listenpreise der Model-Studio-API in CNY je 1 Mio. Token (Stand 31.10.2026, Aktionen ausgenommen): Instruct Eingabe 2, Ausgabe 8 (Peking, Frankfurt, Virginia); Singapur 2,936 bzw. 11,743. Thinking: Eingabe 2, Ausgabe 20 (Singapur 2,936 bzw. 29,357). Ratenlimits 60 Anfragen und 100.000 Token pro Minute. Preise für Drittanbieter und Qwen Chat sind separat zu prüfen; keine Abo- oder Gesamtkostenzusage.

Preisstand 31.10.2026. Offizielle Dokumentation: Qwen3-VL-235B-A22B

Vor dem Einsatz prüfen

- Edition und Modell-ID festhalten: qwen3-vl-235b-a22b-instruct und qwen3-vl-235b-a22b-thinking sind separate Modelle mit eigenen Preisen.
- Kontextbudget nach Zugangsweg prüfen: API-Fenster 131.072 Token gegenüber 256K nativ bei selbst gehosteten Gewichten.
- Preise in CNY und je Region (u.a. Singapur teurer) im Budget erfassen; Ratenlimits von 60 RPM und 100.000 TPM einplanen.

Häufige Fragen zu Qwen3-VL-235B-A22B

Was ist Qwen3-VL-235B-A22B?

Das größte Vision-Language-Modell der Qwen3-VL-Serie von Alibaba Cloud. Es verarbeitet Text, Bilder und Videos zu Text und ist als Instruct- und als Thinking-Edition mit offenen Gewichten erschienen.

Ist Qwen3-VL-235B-A22B Open Source?

Die Modellgewichte beider Editionen stehen unter der Apache License 2.0 bei Hugging Face und ModelScope. Für die gehostete API und Qwen Chat gelten eigene Plattformbedingungen; das ist von der Gewichtslizenz zu trennen.

Wie groß ist das Kontextfenster?

Die Gewichte unterstützen nativ 256K Token, erweiterbar bis 1M. Die Model-Studio-API gibt 131.072 Token Fenster an, mit maximal 129.024 (Instruct) bzw. 126.976 (Thinking) Eingabe-Token und 32.768 Ausgabe-Token. Kontext und Ausgabe sind getrennte Grenzen.

Was kostet Qwen3-VL-235B-A22B über die API?

Laut Model-Studio-Preisliste: Instruct 2 CNY je Million Eingabe-Token und 8 CNY je Million Ausgabe-Token; Thinking 2 CNY Eingabe und 20 CNY Ausgabe. Die Region Singapur ist teurer. Stand: 31.10.2026.

Was unterscheidet Instruct von Thinking?

Gleiche Größe und Architektur, aber eigene Gewichte: Instruct antwortet direkt, Thinking erzeugt eine Denkphase für komplexes Reasoning und ist in der Ausgabe deutlich teurer. In der API sind es zwei verschiedene Modell-IDs.

Kann das Modell Videos verstehen?

Ja, Video ist neben Text und Bild eine dokumentierte Eingabe. Eine Text-Zeitstempel-Ausrichtung soll Ereignisse sekundengenau lokalisieren; der Hersteller nennt stundenlange Videos im erweiterten Kontext als Anwendungsfall.

Alle Antworten beziehen sich auf die verlinkte Herstellerdokumentation und den angegebenen Stand. Kein unabhängiger Benchmark, keine Garantie für Verfügbarkeit oder Ergebnisqualität.

Quellen & Stand

1. Hugging Face: Qwen/Qwen3-VL-235B-A22B-Instruct
   Geprüft: 2026-10-31
2. Hugging Face: Qwen/Qwen3-VL-235B-A22B-Thinking
   Geprüft: 2026-10-31
3. GitHub: QwenLM/Qwen3-VL
   Geprüft: 2026-10-31
4. Alibaba Cloud Community: Qwen3-VL: Sharper Vision, Deeper Thought, Broader Action
   Geprüft: 2026-10-31
5. Alibaba Cloud Model Studio: qwen3-vl-235b-a22b-instruct - Model Info
   Geprüft: 2026-10-31
6. Alibaba Cloud Model Studio: qwen3-vl-235b-a22b-thinking - Model Info
   Geprüft: 2026-10-31

Redaktionelle Zusammenfassung mit KI-Unterstützung, verantwortet von Philipp Paulik. Kein eigener Leistungstest. Änderbare Anbieterangaben gelten für den angegebenen Quellenstand.

Intelligenz Welt · Redaktion: Philipp Paulik · Quellenstand: 2026-10-31
Vollständiges Webportrait: https://intelligenz-welt.de/ki-modelle/qwen3-vl-235b-a22b/
Statischer Export, kein Live-Datenblatt. Preise, Zugriff und Anbieterangaben können sich ändern. Quellen und ihre jeweiligen Abrufstände stehen oben. Impressum · Transparenz